A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments

Wei Pan[1]

Division of Biostatistics, School of Public Health, University of Minnesota

August 3, 2001; Revised October 8, 2001

[1] Division of Biostatistics, A460 Mayo, MMC 303, 420 Delaware Street SE, Minneapolis, MN 55455-0378, USA. Email: weip@biostat.umn.edu, phone: (612)626-2705, and fax: (612)626-0660.

A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments

ABSTRACT

Motivation: A common task in analyzing microarray data is to determine which genes are differentially expressed across two kinds of tissue samples or samples obtained under two experimental conditions. Recently several statistical methods have been proposed to accomplish this goal when there are replicated samples under each condition. However, it may not be clear how these methods compare with each other. Our main goal here is to compare three methods, the t-test, a regression modeling approach (Thomas et al, 2001) and a mixture model approach (Pan et al, 2001) with particular attention to their different modeling assumptions.

Results: It is pointed out that all the three methods are based on using the two-sample t-statistic or its minor variation, but they differ in how to associate a statistical significance level to the corresponding statistic, leading to possibly large difference in the resulting significance levels and the numbers of genes detected. In particular, we give an explicit formula for the test statistic used in the regression approach. Using the leukemia data of Golub et al (1999), we illustrate these points. We also briefly compare the results with those of several other methods, including the empirical Bayesian method of Efron et al (2000) and the Significance Analysis of Microarray (SAM) method of Tusher et al (2001).

Contact: weip@biostat.umn.edu

Key Words: Gene expression; Microarray; Mixture modeling; Regression modeling; t-test.

INTRODUCTION

An exciting development in genomics is the use of microarray technology to simultaneously monitor the expression levels of thousands of genes (or expressed sequence tags) (Brown and Botstein 1999; Lander 1999). A common task is to compare the expression levels of genes in samples drawn from two different tissues or at two different time points or conditions. Specifically, it is of interest to detect genes with differential expression under the two conditions. In early days, the simple method of fold changes was used and now it is known to be unreliable (Chen et al 1997) because statistical variability was not taken account. Since then, many more sophisticated statistical methods have been proposed (e.g. Chen et al 1997; Efron et al 2000; Ideker et al 2000; Newton et al 2001; Tusher et al 2001; Lin et al 2001; Pan et al 2001a). It has also been noticed that data based on a single array may not be reliable and may contain high noises (Lee et al 2000). As the technology advances, microarray experiments are becoming less expensive, which makes the use of multiple arrays (or multiple spots on each array) feasible. In this paper, we consider the detection of differentially expressed genes with replicated measurements of expression levels of each gene under each condition.

A straightforward method is to use the traditional two-sample t-test (e.g. Devore and Peck, 1997). Thomas et al. (2001) proposed a regression modeling approach. Pan et al. (2001a) suggested a mixture model approach, which follows the basic idea of Efron et al (2000) and Tusher et al (2001). However, it is not clear how these methods compare with each other. For practitioners to choose a method, it is important to elucidate various modeling assumptions underlying each method. In this paper, we comparatively review the three methods, the t-test, the regression approach of Thomas et al. (2001), and the mixture model approach of Pan et al. (2001a). In particular, we give an explicit form of the test statistic in the regression method, facilitating the discussion of the connections and differences among the three methods. We apply the three methods to the leukemia data of Golub et al (1999). We also briefly discuss the results of applying the empirical Bayesian (EB) method of Efron et al (2000) and the Significance Analysis of Microarray (SAM) method of Tusher et al (2001) to the same leukemia data.

All the methods are not restricted to any specific microarray technology. From now on, the expression level can refer to a summary measure of relative red to green channel intensities in a fluorescence-labeled cDNA array, a radioactive intensity of a radiolabeled cDNA array, or a summary difference of the perfect match (PM) and mis-match (MM) scores from an oligonucleotide array. The gene expression levels may have been suitably preprocessed, including dimension reduction, data normalization and data transformation (e.g. Dudoit et al 2000; Efron et al 2000; Li and Wong 2001; Kerr et al 2000; Yang et al 2000). To focus on the main issue, we assume that all the methods use the data preprocessed in the same way.

METHODS

Data

Suppose that Y_{jk} is the expression level of gene j in array k ($j = 1, ..., n$; $k = 1, ..., K_1, K_1 + 1, ..., K1 + K2$). Suppose that the first $K1$ and last K_2 arrays are obtained under the two conditions respectively.

A general statistical model is

$$Y_{jk} = a_j + b_j x_k + \epsilon_{jk}, \tag{1}$$

where $x_k = 1$ for $1 \leq k \leq K_1$ and $x_k = 0$ for $K_1 + 1 \leq k \leq K_1 + K_2$, and ϵ_{jk} are random errors with mean 0. Hence $a_j + b_j$ and a_j are the mean expression levels of gene j under the two conditions respectively. Determining whether a gene has differential expression is equivalent to testing for the null hypothesis

$$H_0\text{: } b_j = 0 \text{ against } H_1\text{:}b_j \neq 0.$$

A statistical test consists of two parts. The first is to construct a summary test statistic. The second is to determine the significance level or p-value associated with the test statistic. The p-value is usually calculated based on the null distribution of the test statistic (i.e. the distribution of the test statistic under H_0), which may be specified or estimated via modeling assumptions.

To focus on the main issue, we use $\alpha = 0.01$ as the genome-wide significance level, and we use the Bonferroni adjustment to deal with multiple comparisons. Hence the test- or gene-specific significance level (for a two-sided test) is $\alpha^* = \alpha/(2n)$. We do not consider other possibly better adjustment methods for multiple comparisons (e.g. Dudoit et al 2000; Thomas et al 2001). Most, if not all, statistical tests can be modified accordingly for a multiple comparison adjustment.

In the following, we review the three methods along the line.

The t-test

There are several versions of the two-sample t-test, depending on whether the sample size (i.e. K_1 and K_2) is large and whether it is reasonable to assume that gene expression levels have an equal variance under the two conditions (Devore and Peck, 1997, Sections 10.1-10.2). Because usually both K_1 and K_2 are small, and there is evidence to support unequal variances (e.g. Thomas et al, 2001), we will only discuss the t-test with two independent small Normal samples with unequal variances.

Let the sample means and variances of Y_{jk}'s for gene j under the two conditions be

$$\bar{Y}_{j(1)} = \frac{\sum_{k=1}^{K_1} Y_{jk}}{K_1}, \quad \bar{Y}_{j(2)} = \frac{\sum_{k=K_1+1}^{K_1+K_2} Y_{jk}}{K_2}$$

and

$$s^2_{j(1)} = \frac{\sum_{k=1}^{K_1}(Y_{jk} - \bar{Y}_{j(1)})^2}{K_1 - 1}, \quad s^2_{j(2)} = \frac{\sum_{k=K_1+1}^{K_1+K_2}(Y_{jk} - \bar{Y}_{j(2)})^2}{K_2 - 1}.$$

The t-statistic is

$$Z_j = \frac{\bar{Y}_{j(1)} - \bar{Y}_{j(2)}}{\sqrt{s^2_{j(1)}/K_1 + s^2_{j(2)}/K_2}}. \tag{2}$$

Under the Normality assumption for Y_{jk}, Z_j approximately has a t-distribution with degrees of freedom

$$d_j = \frac{(s^2_{j(1)}/K_1 + s^2_{j(2)}/K_2)^2}{(s^2_{j(1)}/K_1)^2/(K_1 - 1) + (s^2_{j(2)}/K_2)^2/(K_2 - 1)}.$$

This t-test was proposed by Welch (1937). Its method of calculating the degrees of freedom is similar to the idea of the Satterthwaite approximation. The good performance of Welch's t-test, compared with many other alternatives (e.g. approximating d_j by $K_1 + K_2 - 2$), has been well documented (e.g. Scheffe 1970; Best and Rayner 1987).

We will use the Welch t-test throughout. Note that the Welch t-test is specifically designed to handle the possibility of having unequal variances. If we ignore the part of multiple comparison adjustment, Dudoit et al (2000) adopt the same t-statistic, but calculate the p-value by permutation.

A regression modeling approach

Thomas et al (2001) proposed a regression modeling approach. In their original formulation, data preprocessing and testing for differential expression are coupled together. To focus on the main issue, we only consider the testing part of their approach throughout the paper. Then their model is the same as (1) (after ignoring the preprocessing step). Treating (1) as a regression model, they proposed to estimate (a_j, b_j) using the weighted least square method, and then to estimate the variance of $\hat{b}_j$ using the robust or sandwich variance estimator, say $Var(\hat{b}_j)$. Their test statistic is

$$Z_j' = \hat{b}_j / \sqrt{Var(\hat{b}_j)},$$

with reference to the (asymptotically) Normal distribution.

Thomas et al observed that the result based on using Z_j' is close to that of using the t-statistic Z_j. But because they did not give an explicit formula for Z_j', a theoretical explanation was unavailable. Next we give an explicit formula for Z_j', which will shed light on the nature of Z_j'. Note that (1) can be formulated as a linear regression problem. If we denote the corresponding design matrix as X, it is easy to verify that $X'X$ is block-diagonal. Hence, the least-squares estimates of (a_j, b_j) are independent with each other for different genes. In fact, it is easy to verify that the least-squares estimate of b_j is

$$\hat{b}_j = \bar{Y}_{j(1)} - \bar{Y}_{j(2)}.$$

Then, as pointed out by Drum and McCullagh (1993), the robust variance estimator of $\hat{b}_j$ is

$$Var(\hat{b}_j) = \frac{s_{j(1)}^2}{K_1} \frac{K_1 - 1}{K_1} + \frac{s_{j(2)}^2}{K_2} \frac{K_2 - 1}{K_2},$$

which can be also verified directly.

Hence, it can be seen that the statistic Z_j' of Thomas et al has a similar form to the usual t-statistic Z_j with the minor difference in how to estimate the variances: rather than the unbiased sample variances as used in Z_j, the maximum likelihood estimator of the variance (under the Normality assumption for Y_{jk}'s) is used in Z_j'. It is obvious that Z_j' and Z_j are equivalent as both K_1 and K_2 tend to the infinity. However, for small K_1 and K_2, Z_j is preferred due to the unbiasedness of its variance estimator involved.

Furthermore, using a standard Normal distribution to calculate the p-value for Z_j' is based on the assumption that both K_1 and K_2 are large, which however does not hold in many microarray experiments. Therefore, as acknowledged by Thomas et al, the Normality assumption for Z_j' may be too strong and may not work well in practice.

Note that the use of the sandwich variance estimator proposed by Thomas et al (2001) is novel. It works asymptotically even if random errors ϵ_{jk}'s in (1) have different variances for different j, or even for the same gene j under the two conditions. More often, for a linear regression model like (1), it is assumed that ϵ_{jk}'s are a random sample (iid) from the same distribution, and hence have the same variance. If we ignore the preprocessing that corrects array and other effects, the latter assumption on iid ϵ_{jk}'s is taken by Kerr et al (2000).

A mixture modeling approach

A common problem with the above t-test and the regression approach is their strong assumptions on the null distributions of the test statistics. In contrast, following the idea of Efron et al (2000) and Tusher et al (2001), Pan et al (2001a) proposed to estimate the null distribution directly. The method takes full advantage of the existence of replicated data, but it does require that both K_1 and K_2 are even numbers.

A key step is to construct the following null statistics:

$$z_j = \frac{Y_{j(1)}p_j/K_1 - Y_{j(2)}q_j/K_2}{\sqrt{s^2_{j(1)}/K_1 + s^2_{j(2)}/K_2}}, \qquad (3)$$

where $Y_{j(1)} = (Y_{j1}, ..., Y_{jK_1})$, $Y_{j(2)} = (Y_{j,K_1+1}, ..., Y_{j,K_1+K_2})$, p_j is a random permutation of a column vector containing $K_1/2$ 1's and -1's respectively, and q_j is a random permutation of a column vector containing $K_2/2$ 1's and -1's respectively. Suppose that the distribution density functions of z_j and Z_j are respectively f_0 and f. Under the weak assumption that the random errors ϵ_{jk} have a distribution symmetric about its mean 0, then under H_0, z_j and Z_j have the same distribution $f_0 = f$.

Using z_j's and Z_j's we can estimate the distributions f_0 and f respectively. We will discuss how to estimate f_0 and f later. For the moment, suppose f_0 and f are known (or more precisely, taken at their estimates). For any given Z, we can use the likelihood ratio test statistic

$$LR(Z) = f_0(Z)/f(Z)$$

to test for H_0. A small value of $LR(Z)$, say $LR(Z) < c$, provides evidence to reject H_0. The cut-off point c is determined such that the type I error rate is

$$\frac{\alpha}{n} = \int_{LR(z)<c} f_0(z)dz, \qquad (4)$$

where α is the genome-wide significance level.

Now we discuss how to estimate f_0 and f. Pan et al proposed using a Normal mixture to model each distribution:

$$f_0(z; \Omega_{g_0}) = \sum_{i=1}^{g_0} \pi_i \phi(z; \mu_i, V_i),$$

where $\phi(.; \mu_i, V_i)$ denotes the normal density function with mean μ_i and variance V_i, and π_i's are mixing proportions. Ω_{g_0} represents all unknown parameters $\{(\pi_i, \mu_i, V_i) : i = 1, ...g_0\}$ in a g_0-component mixture model. The number of components can be selected adaptively. Similarly, a Normal mixture model can be used for f.

The Normal mixture model is flexible and powerful, and is widely used in many applications. It is usually fitted by maximum likelihood using the expectation-maximization (EM) algorithm (Dempster et al. 1977). To determine the number of components g_0, we can use various model selection criteria, of which the Bayesian Information Criterion (BIC) (Schwartz 1978) is favored in some empirical studies (Fraley and Raftery 1998):

$$BIC = -2 \log L(\hat{\Omega}_{g_0}) + \nu_{g_0} \log(N),$$

where $L(\hat{\Omega}_{g_0})$ is the maximized likelihood function and $\nu_{g_0} = 3g_0 - 1$ is the number of independent parameters in Ω_{g_0}. In using the BIC, one first fits a series of models with various values of g_0, then picks up the g_0 corresponding to the first local minimum of BIC (Fraley and Raftery 1998).

We used the EMMIX, a stand-alone Fortran program for fitting a Normal mixture model using maximum likelihood method. It was implemented by McLachlan et al. (1999) and is freely available from the web at

http://www.maths.uq.oz.au/∼gjm/emmix/emmix.html.

It has many interesting features, including multiple starts of the EM algorithm and calculation of model selection criteria.

RESULTS

Data

We apply the methods to the leukemia data of Golub et al (1999), which consists of 27 acute lymphoblastic leukemia (ALL) samples and 11 acute myeloid leukemia (AML) samples. The goal is to find genes with differential expression between ALL and AML. Thomas et al also analyzed this data set. Since the mixture model approach requires even numbers of samples, we randomly take the first $K_1 = 26$ ALL samples and last $K_2 = 10$ AML samples for use. There are $n = 7129$ genes in each sample. As mentioned earlier, we take the genome-wide significance level at the usual $\alpha = 0.01$ level, and use the Bonferroni method to adjust for multiple comparisons. Hence the gene-specific significance level being used is $\alpha^* = 0.01/(7129 * 2) = 7.014 \times 10^{-7}$.

Data preprocessing is accomplished for each sample by subtracting its median and dividing by its quartile range (the difference between the first and the third quartiles). Note that rather than the commonly used mean and standard deviation, we used the median and quartile range because the latter two are more robust estimators for the center and the dispersion of a distribution respectively. All the methods are applied to thus preprocessed data.

Fitted mixture models

We fitted three mixture models for either f_0 or f with 1 to 3 components (Table 1). Based on BIC, we can see that a two-component mixture is the best for each and the fitted models are:

$$f_0(z) = 0.479\phi(z; -0.746, 0.697) + 0.521\phi(z; 0.739, 0.641),$$

and

$$f(z) = 0.518\phi(z; -0.318, 1.803) + 0.482\phi(z; 0.781, 4.501).$$

Figure 1 presents the histograms and the fitted models. For comparison, the density function of a t-distribution with degrees of freedom 34 is imposed in Figure 1(a). It can be seen that the t distribution has much heavier tails than the estimated f_0. For Figure 1(b), we present both the estimated f_0 and f. It can be seen that, unsurprisingly, f has heavier tails than f_0.

Table 1 about here

Figure 1 about here

The LR function is depicted in Figure 2. Using the bisection method (Press et al 1992, p.353) to solve equation (4), we obtain the cut-off point $c = 0.0003437$, yielding a rejection region of $\{Z : Z < -4.8877 \text{ or } Z > 4.4019\}$ for H_0.

Figure 2 about here

Total number of genes with differential expression

The test statistics Z_j and Z'_j are easy to calculate. But the three methods have different rejection regions for H_0.

For t-test, because the degrees of freedom d_j vary with j, the rejection region is gene-specific: $\{Z : |Z| > t(\alpha^*, d_j)\}$, where $t(\alpha^*, d_j)$ is the upper α^*-percentile for a t-distribution with degrees of freedom d_j. The distribution of n calculated d_j's ranges from 9 to 34 with mean and median 17.8 and 19.6 respectively. Note that the smaller the d_j, the large the $t(\alpha^*, d_j)$. To give a rough idea of the rejection region, we consider the conservative situation $d_j = 34$: the rejection region is $\{Z : |Z| > 5.8369\}$.

For the regression modeling method, since the null distribution is assumed to be a standard Normal, the rejection region is $\{Z' : |Z'| > 4.8246\}$, where 4.8246 is the upper α^*-percentile for a standard Normal distribution.

As described earlier, the rejection region for the mixture model method is $\{Z : Z < -4.8877 \text{ or } Z > 4.4019\}$.

Comparing the rejection regions of the three methods and the estimated null distribution and a t distribution in Figure 1(a), we can see that the t-test is probably too conservative. Unsurprisingly, the t-test detects total 20 genes with significant expression changes, whereas the regression method and the mixture model method find 157 and 187 genes respectively.

Thomas et al provided some biological justifications for many identified genes.

Top genes with differential expression

Tables 2 and 3 list top 25 genes which are more highly expressed in AML and in ALL respectively. In general, the regression and the mixture model methods give very similar rankings; This can be explained by the closeness of Z_j to Z'_j. In particular, it is reassuring that the two spots of the same gene, TCF3 (E2A), in Table 3 are ranked closedly as number 14 and number 16 by both methods. The results of the three methods are in good agreement in Table 3. However, only the top 6 genes in Table 2 are identified as such by all the three methods. This may be due to the fact that the absolute values of the test statistics in Table 3 are much larger than those in Table 2. In other words, there is stronger evidence to suggest differential expression for the top genes more highly expressed in ALL than that for those more highly expressed in AML. In fact, according to the mixture model method, only the top 16 genes in Table 2 are identified as having significant expression change.

Table 2 about here

Table 3 about here

Note that the ranking of the t-test is based on the corresponding p-values. Since the degrees of freedom of the null distribution in the t-test are gene-specific, the resulting ranking is different from that based on the test statistics Z_j. Thomas et al (2001) reported a good agreement between the t-test and the regression approach. The reason is that they used a fixed number (36) for the degrees of freedom in the t-test (LP Zhao, personal communications). If we do it that way, then the same conclusion can be drawn. Also, since the null distributions for the regression method and the mixture model method are fixed (i.e. non-gene-specific), the ranking based on Z_j or Z'_j should be the same as that based on the corresponding p-values.

For the purpose of comparison, we also give the ranking results taken from Thomas et al (2001). Most of the genes listed are also reported by them. However, the specific ranking may be very different. This may be largely due to the different methods used in preprocessing the data. This demonstrates the importance of data preprocessing.

DISCUSSION

A comparative summary of the three methods

We have given an explicit expression of the test statistic Z'_j for the regression approach of Thomas et al, from which we can see that it has a similar form to the t-statistic, which is also used in the mixture model approach. Hence, the three methods usually give similar results in terms of the test statistics. However, they differ in how to determine the statistical significance level (or rejection region). For small sample sizes, both the t-test and the regression approach depend on the strong parametric assumptions, the t-distribution of Z_j (or equivalently, the Normality assumption on the random errors) and the Normal distribution of Z'_j respectively. It is possible that these parametric assumptions are violated in practice when small sample sizes are more common, though the two methods are asymptotically valid (with large sample sizes). In contrast, the mixture model approach estimates the null distribution directly. It takes advantage of the existence of multiple samples to construct the null scores z_j's, and the large number of genes makes it feasible to estimate the null distribution f_0 (and f) nonparametrically. Note that the null distribution f_0 is for random errors, not for the gene expression levels. Of course, the mixture model method (as the EB method of Efron et al and SAM) also has its own modeling assumptions: it is assumed that the random errors have symmetric distributions, and after a suitable standardization (here we divide them by the sample variances), the random errors from all the genes have a common distribution. We believe that these assumptions are weak and reasonable. In particular, they are weaker than the Normality assumption used in the t-test.

Note that the Normality assumption in the t-test is required only for small K_1 and K_2. If both K_1 and K_2 tend to the infinity, d_j also goes to the infinity, implying that the null distribution reduces to a standard Normal. Thus, if both K_1 and K_2 are large, such as > 30 as suggested by many introductory statistics textbooks (e.g. Devore and Peck, 1997, p.352), one can use a standard Normal as the null distribution for the t-test. The asymptotic Normality assumption for the null distribution of Z'_j in the regression approach also requires large K_1 and K_2. Hence, with large K_1 and K_2, both methods are essentially nonparametric.

With practical numbers of samples (i.e small K_1 and K_2), however, the power of the t-test is limited (due to the too small degrees of freedom), whereas the Normality assumption for the regression approach is more likely to be seriously violated. These are the situations where the mixture model approach and other similar approaches (Efron et al 2000, Tusher et al 2001) are more attractive. An advantage of the regression approach is its flexibility: it can be extended to model more complex biological processes (Zhao et al., 2001). An attractive point of the mixture model approach is its use for sample size/power calculations (Pan et al 2001b).

A brief comparison with other approaches

The Wilcoxon rank sum test (equivalent to Mann-Whitney test) has also been used as an alternative to the t-test in two-sample comparisons with microarray data. Because it is nonparametric, it avoids the possibly questionable parametric assumption used in the t-test. However, as demonstrated by Thomas et al., the price we pay for the robustness of the Wilcoxon test is the loss of power: when applied to the leukemia data, it does not find any gene with significant expression change. This is also related to another often neglected issue: the Wilcoxon test requires that the two samples have distribution functions with the same shape (with the only difference in their location parameters). This implies, strictly speaking, that it is not applicable if the expression levels of a gene may have unequal variances under the two conditions, which is exactly the same reason why we prefer the t-test with unequal variances to that with an equal variance. These same issues remain with the use of other permutation-based nonparametric tests.

Figure 3 about here

The mixture model approach follows the novel idea of the EB approach of Efron et al (2000) and of the SAM of Tusher et al (2001): estimating the null distribution using z_j's. They belong to the same family with the same basic modeling assumptions. Here, we give a brief comparison of these methods by applying them to the same leukemia data. The general conclusion below is similar to that of Pan et al (2001a). To save space, we do not go to the details of the EB and SAM methods; The reader is referred to the above references for more details.

For the EB approach, Efron et al (2000) proposed a logistic regression method to estimate the likelihood ratio statistic $LR(Z)$ (Figure 3(a)), which is close to that obtained by the mixture method (Figure 2). Using $LR(Z)$, Efron et al derived a lower bound of the posterior probability that a gene with the t-statistic Z_j has differential expression, $Pr(Event|Z_j)$. The posterior probability is drawn as a function of Z in Figure 3 (b). The qualitative conclusion is the same as other methods: as $|Z_j|$ increases, there is stronger evidence to reject H_0. Corresponding to the rejection region by the mixture model method, the estimated posterior probabilities are $Pr(Event|-4.9) = 0.978$ and $Pr(Event|4.4) = 0.982$. The posterior probability is closely related to the so-called false discovery rate (FDR) (Efron et al, 2001). The FDR is used as an alternative to controlling the false positive rate (i.e. Type I error rate) in handling multiple comparisons (Benjamini and Hochberg 1995). A potential problem is that, since only the lower bound of the posterior probability is actually estimated and given, the interpretation of the result in terms of significance level may be conservative. Nontheless, there are many interesting features in the EB approach.

Figure 4 about here

As pointed out by Efron et al, SAM is best suited to detecting a small number of genes with differential expression, which is not true for the leukemia data. Pan et al (2001a) also pointed out a problem with SAM under these situations. Although the SAM was originally designed to control the FDR, if desired, it can be also directly applied to control the Type I error. Using $B = 20$ versions of random permutations of z_j scores, one can calculate the expected order statistics of z_j's,

$\bar{z}_{(j)}$. If we use $s = 2$, the estimated false positives and true positives are 0.1 and 297 respectively. If we use $s = 2.1$, the estimated false positives and true positives are 0 and 267 respectively. Figure 4 presents the results of SAM using $s = 2$, where the identified 297 genes with differential expression are those satisfying $|Z_{(j)} - \bar{z}_{(j)}| > s$ with $Z_{(j)}$'s being the order statistics of Z_j's. Since simulation is used in SAM, it is in general difficult to obtain results for a given Type I error α^*. On the other hand, there are many attractive points of SAM. For instance, it does not have strong parametric assumptions and does not involve any complex estimation procedures (i.e. only order statistics are involved). In particular, it compares Z_j's of all the genes *collectively* with their z_j's. In contrast, all the other methods test gene by gene, which may be less efficient.

ACKNOWLEDGMENTS

The author thanks Drs Brad Efron, Rob Tibshirani and Lue Ping Zhao for helpful communications. In particular, Dr Zhao read an earlier version. The author is also grateful to the referees and editor for constructive comments.

REFERENCES

1. Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289-300.

2. Best, D.J. and Rayner, J.C.W. (1987). Welch's approximate solution for the Behrens-Fisher problem. *Technometrics*, **29**, 205-210.

3. Brown, P and Botstein, D.. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics (Suppl.)*, **21**, 33-37.

4. Chen Y, Dougherty ER and Bittner ML (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J Biomedical Optics*, 2, 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*, **39**, 1-38.

6. Devore, J. and Peck, R. (1997). *Statistics: the Exploration and Analysis of Data.* 3rd ed. Duxbury Press: Pacific Grove, CA.

7. Drum, M. and McCullagh, P. (1993). Comment on "Regression models for discrete longitudinal responses". *Statistical Science*, **8**, 300-301.

8. Dudoit S, Yang YH, Callow MJ and Speed TP (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Tech Rept, Stat Dept, UC-Berkeley, 2000.

9. Efron, B., Tibshirani, R., Goss, V. and Chu, G. (2000). Microarrays and Their Use in a Comparative Experiment. Manuscript.
Available at `http://www-stat.stanford.edu/~tibs/research.html`.

10. Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. To appear in *Journal of the American Statistical Association.*

11. Fraley, C. and Raftery, A.E. (1998). How many clusters? Which clustering methods? – Answers via model-based cluster analysis. *The Computer Journal*, **41**, 578-588.

12. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **285**, 531-537.

13. Ideker, T., Thorsson, V., Siehel, A.F. and Hood, L.E. (2000). Testing for differentially-expressed genes by maximum likelihood analysis of microarray data. *Journal of Computational Biology*, **7**, 805-817.

14. Kerr, M.K., Martin, M. and Churchill, G.A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, **7**, 819-837.

15. Lander, E.S. (1999). Array of hope. *Nature Genetics (Suppl.)*, **21**, 3-4.

16. Lee, M-L T., Kuo, F.C., Whitmore, G.A. and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Nat. Acad. Sci*, **97**, 9834-9839.

17. Li, C. and Wong, W.H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *PNAS*, **98**, 31-36.

18. Lin, Y., Nadler, S. T., Attie, A. D., and Yandell, B. S. (2001). Mining for Low-abundance Transcripts in Microarray Data. Manuscript.

Available at http://www.stat.wisc.edu/∼yilin/.

19. McLachlan, G.L., Peel, D., Basford, K.E. and Adams, P. (1999). Fitting of mixtures of normal and t-components. *Journal of Statistical Software*, **4**. (http://www.stat.ucla.edu/journals/jss)

20. 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.

21. Pan, W., Lin, J. and Le, C. (2001a). A Mixture Model Approach to Detecting Differentially Expressed Genes with Microarray Data. Tech Rept, Division of Biostatistics, U of Minnesota. Available at http://www.biostat.umn.edu/cgi-bin/rrs?print+2001.

22. Pan, W., Lin, J. and Le, C. (2001b). How Many Replicates of Arrays Are Required to Detect Gene Expression Changes in Microarray Experiments? A Mixture Model Approach. Tech Rept, Division of Biostatistics, U of Minnesota.
Available at http://www.biostat.umn.edu/cgi-bin/rrs?print+2001.

23. 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. Cambridge: New York.

24. Scheffe, H. (1970). Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Association*, **65**, 1501-1508.

25. Schwartz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics*, **6**, 461-464.

26. Thomas, J.G., Olson, J.M., Tapscott, S.J. and Zhao, L.P. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research*, **11**, 1227-1236.

27. Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley, New York.

28. Tusher, V.G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Nat. Acad. Sci*, **98**, 5116-5121.

29. Welch, B.L. (1947). The generalization of 'Students' problem when several different population variances are involved. *Biometrika*, **34**, 28-35.

30. Yang, Y.H., Buckley, M.J., Dudoit, S. and Speed, T.P. (2000). Comparison of methods for image analysis on cDNA microarray data. Tech Rept, Stat Dept, UC-Berkeley, 2000.

31. Zhao, L.P., Prentice, R. and Breeden, L. (2001). Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc National Acedemy of Sciences USA*, **98**, 5631-5636.

Table 1: BIC for fitted mixture models with various numbers of components g.

g	1	2	3
f_0	21656.7	21585.4	21598.2
f	28986.1	28833.3	28857.1

Table 2: Top 25 genes more highly expressed in AML than in ALL.

				t	Regression	Thomas et al
Gene	Probe	Z_j	Z'_j	rank	rank	rank
FAH	M55150	-7.875	-8.248	1	1	1
Neuromedin B	M21551	-5.767	-5.963	2	2	2
CDC25A	M81933	-5.540	-5.778	6	3	4
NAB50	U63289	-5.508	-5.724	3	4	14
FTL	M11147	-5.426	-5.638	4	5	12
Metargidin	U41767	-5.394	-5.617	5	6	10
LYN	M16038	-5.284	-5.563	-	7	7
LTC4S	U50136	-5.264	-5.541	-	8	3
Metallothionein I-B	M13485	-5.096	-5.343	17	9	-
IGIF	D49950	-5.042	-5.296	23	10	8
PPlase	M80254	-5.023	-5.281	-	11	21
Inositol 1,3,4-trisphosphate 5/6-kinase	U51336	-5.001	-5.216	10	13	-
Zyxin	X95735	-4.976	-5.242	-	12	6
ATP6C	M62762	-4.933	-5.174	22	14	9
CMKBR7	L08177	-4.894	-5.139	-	15	-
Chloride channel (putative) 2163bp	Z30644	-4.884	-5.089	11	16	19
Thrombospondin 1	U12471	-4.847	-5.051	14	17	5
MDU1 Antigen	M21904	-4.795	-4.996	16	19	-
Proto-oncogene BCL3	U05681	-4.783	-5.002	18	18	-
GST-II	U77604	-4.765	-4.918	7	20	-
Calnexin	D50310	-4.723	-4.914	13	21	15
Polyadenylate binding protein II	Z48501	-4.712	-4.895	9	22	17
Sodium channel protein	M81758	-4.628	-4.789	8	24	-
HoxA9	U82759	-4.592	-4.824	-	23	13
PLCB2	M95678	-4.558	-4.779	-	25	16

Table 3: Top 25 genes more highly expressed in ALL than in AML.

Gene	Probe	Z_j	Z'_j	t rank	Regression rank	Thomas et al rank
P48	X74262	8.083	8.286	1	1	2
ACADM	M91432	7.113	7.278	3	2	11
Macmarks	HG1612-HT1612	6.945	7.149	2	4	5
MYL1	M31211	6.929	7.151	4	3	4
Proteasome iota chain	X59417	6.722	6.896	5	5	3
Adenosine triphosphatase, calcium	Z69881	6.683	6.830	8	6	17
C-myb	U22376	6.587	6.742	7	7	1
IEF SSP 9502	L07758	6.535	6.742	6	8	18
Inducible protein	L47738	6.467	6.624	9	9	7
hdlc1	U32944	6.394	6.532	13	11	20
DHPS	U26266	6.364	6.573	10	10	-
Cyclin D3	M92287	6.303	6.457	11	12	14
MB-1	U05259	6.250	6.378	18	13	8
TCF3 (E2A)	M65214	6.169	6.355	12	14	6
CRYZ	L13278	6.047	6.222	14	15	9
TCF3 (E2A)	M31523	6.007	6.141	21	16	16
Nucleolysin TIA-1	M77142	5.986	6.109	25	19	-
Thymopoietin beta	U09087	5.968	6.124	15	18	13
MCM3	D38073	5.948	6.130	16	17	19
SRB	U38846	5.928	6.103	17	20	-
SPTAN1	J05243	5.868	6.026	20	23	22
Transcriptional activator hSNF2b	D26156	5.859	6.027	19	22	10
ALDR1	X15414	5.822	6.071	-	21	21
HKR-T1	S50223	5.807	6.013	-	24	25
T-complex protein 1, gamma subunit	X74801	5.743	5.878	-	-	-

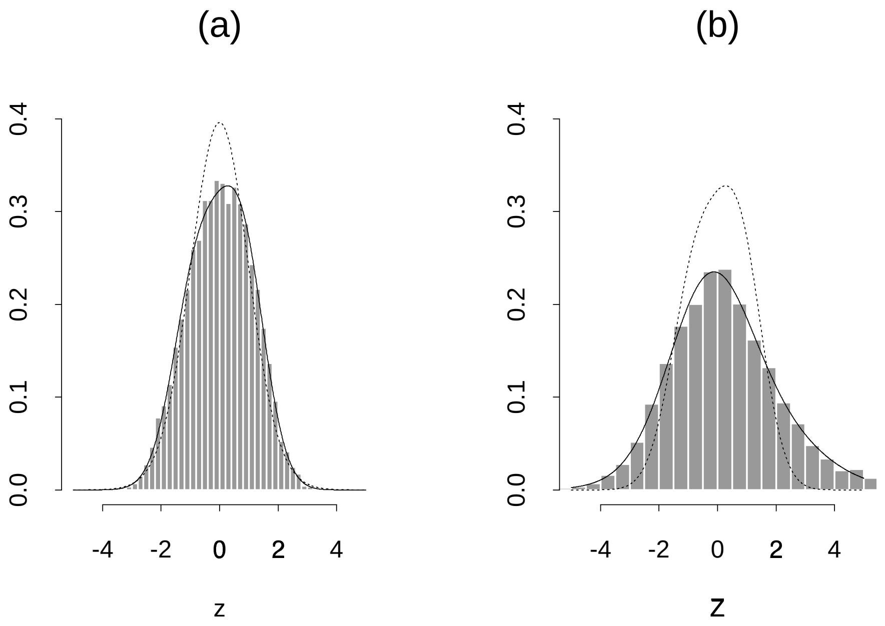

Figure 1: Histograms and fitted mixture models (solid lines) for z_j in (a) and for Z_j in (b). In (a) the dotted line is a t-distribution with 34 degrees of freedom. In (b) the dotted line is the fitted f_0.

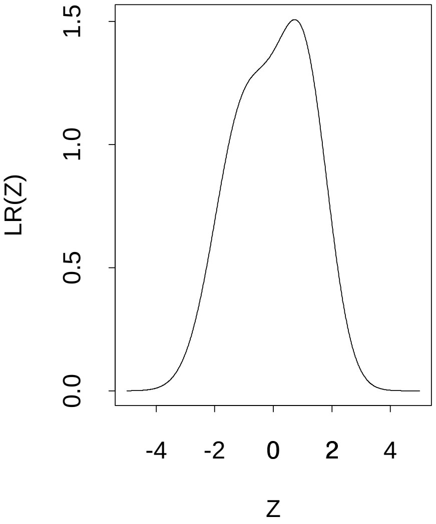

Figure 2: The likelihood ratio function.

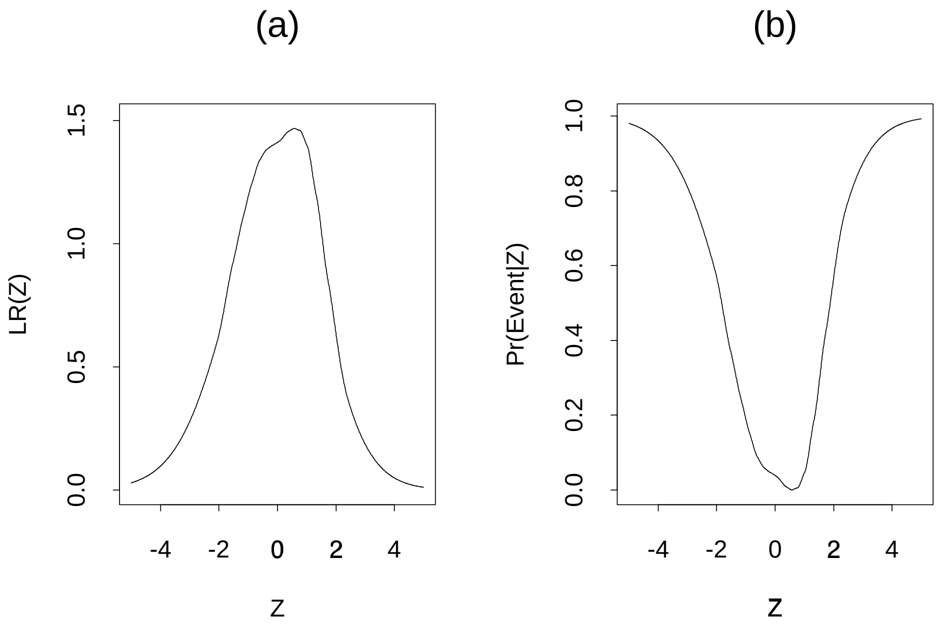

Figure 3: The results of the empirical Bayesian method: (a) the likelihood ratio function using the logistic regression method; (b) the estimated posterior probability function.

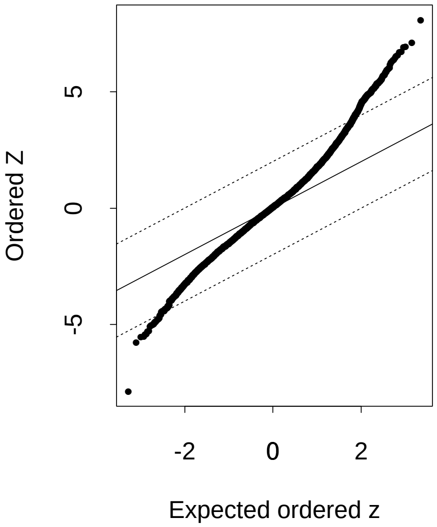

Figure 4: The results of SAM. Any gene corresponding to a point out of the bounds of the dotted lines is interpreted as having a significant expression change.