Analysis of Microarray Gene Expression Data Using a Mixture Model

Al Bartolucci[1], David B. Allison [1], Sejong Bae[2], and Karan P. Singh [2]

[1] Department of Biostatistics, University of Alabama at Birmingham, USA
[2] Department of Biostatistics, University of North Texas Health Science Center, USA
Email: sbae@hsc.unt.edu

Keywords: *Microarray Gene Data, Mixture Model, comparison*

EXTENDED ABSTRACT

The analysis of microarray data remains a challenge as one wish to investigate the possibility of the expression of thousands of genes across multiple samples. Naturally the issue of multiplicity arises as one examines the significance of large numbers of genes. Recently, one of the coauthors, DBA, and colleagues developed a mixed model approach to this very problem with successful application to a mouse data model. In this particular setting one circumvents the false positive issue using a mixture distribution of the p-values. Simultaneously one addresses several issues such as 1) whether we have any statistically significant evidence in any of the genes, 2) what is the best estimate of the number of genes in which there is a true difference in gene expression?, 3) is there a threshold which signals a criteria above which genes should be investigated further?, and 4) what is the possible proportion of false negatives in those genes declared "not interesting" ? The objective of this study was to investigate this procedure further and illustrate its usefulness and relevance in the current work on microarray data analysis.

1. INTRODUCTION

In microarray data analysis we want to examine if certain genes have difference in expression. e.g. disease vs. no disease, characteristic vs. not having characteristic, condition vs. no condition, etc**.** Of those genes determined to be expressed, what proportion are likely to be false leads? Challenges involved in analyzing microarray data include sample size or number of cases is small (humans, mice, other species), however, the number of genes or probes is large (hundreds or thousands), multiplicity issues occurring from numerous comparisons. Mixture Model approach was proposed by Allison et. al. (2002). Many statistical tests are conducted from which one obtains a distribution of p-values and there is information in the p-values that can be exploited. In this paper, we demonstrate the mixture model approach by applying it to p-values generated by several statistical tests comparing two means from several hundred probes. We further examine it's behavior under different assumptions (equality, non equality of variances) and approaches (permutation, bootstrapping).

2. METHODS

We made following assumptions: assume independence of gene expression levels across genes. Assume N=2n cases divided evenly into two groups of n cases each (n=same for each group not a requirement)

H0 : No difference in gene expression between two groups for the ith gene., i=1,…,k
Under H0 the distribution of p-values is uniform on [0,1].

H1 : Alternative is observed distribution of p-values is significantly different from a uniform distribution.
Parker and Rothenberg (1988) point out that any distribution on [0,1] can be modeled as a mixture of V separate component distributions where the jth component (j=1,….,V) is a beta distribution with parameters, r_j and s_j.
The pdf of the beta distribution is
$\beta(r,s)(x) = f(x|r,s) = [xr\text{-}1(1\text{-}x)s\text{-}1]/B(r,s)$
where $B(r,s)=\int[0,1]\ u\ r\text{-}1(1\text{-}u)s\text{-}1du \quad r>0,\ s>0.$
Note the uniform on [0,1] is the special case where r=s=1.
The log likelihood for the collection of k p-values from a model with ν+1 components is:
$L_{V+1} = \sum_i\text{=1,n } \ln[\lambda 0\beta(1,1)(xi) + \sum j\text{=1,}\nu\ \lambda_j\beta(rj,sj)(xi)]$
where x_i = p-value for the $_i$th test.
λ_0 = probability a randomly chosen test from the collection of tests is for a gene for which there is no population difference in gene expression (i.e. test of a true null hypothesis)
λ_j = probability a randomly chosen test from the collection of tests is for a gene from the jth component distribution which yields a true population difference in gene expression (i.e. test of a false null hypothesis)

One can obtain mle's of λ_j ,s_j , and r_j iteratively for the log expression, L_{V+1}, subject to the constraint,

$1=\lambda_0+ \Sigma j=1,\nu\ \lambda_j$ and $0\leq \lambda_j \leq 1$ for all λ_j.

-One can test for ν components by computing the statistics,

$$Q=(L_\nu - L_{\nu-1})$$

(Note this does not have a chi square distribution with 3 df, but can be computed using bootstrap, See Allison et. al, 2002).

The best estimate for the number of genes for which there is a true difference in gene expression is simply,

$k(1- \text{mle}\lambda_0)$

where mleλ_0 is the ml estimate of λ_0.

One can compute the 100(1-α)% confidence interval around mleλ_0 by usual bootstrap.

For a particular p-value of interest one can compute:
PTP = posterior true positives
= the proportion of genes with a true differential expression among the genes which are declared interesting via p-value ≤ x.
PFP = posterior false positive
= proportion of genes with no differential expression among the genes which are declared interesting via p-value ≤ x.

For a particular p-value of interest one can also compute:
PFN=posterior false negatives
=the proportion of genes with a true differential expression among the genes which are declared not interesting via p-value > x.
EDR =expected discovery rate
= expected proportion of genes that are truly differentially expressed that will be declared to be differentially expressed

3. RESULTS

We examined obesity data for difference in gene expression of obese (n=19) vs. non obese (n=20) for 300 gene probes. A p-value was generated for each probe of obese vs. non obese by four tests of means:
Permutation* (equal variance t-test assumption)
Permutation (unequal variance t-test)
Pivotal bootstrap (equal variance t-test)
Pivotal bootstrap (unequal variance t-test)
*See Brand, J et al (2006) for explanation of these tests as related to gene expression data.

Figure 1. Permutation-Unequal Variance (PUV)

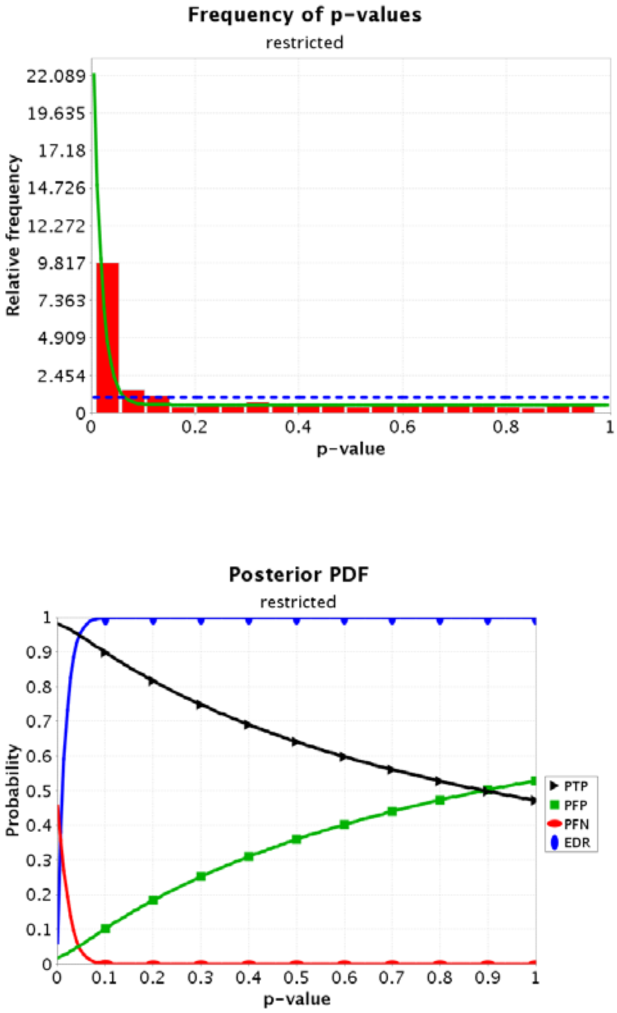
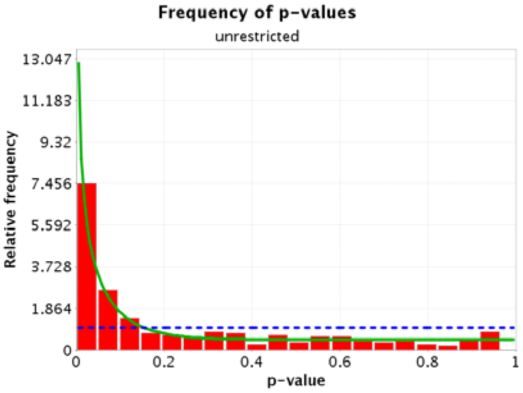
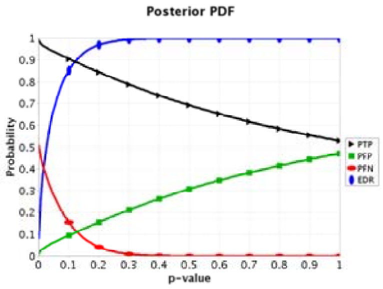


Figure 2. Bootstrap Unequal Variance (BUV)



Bootstrap Unequal Variance (BUV)



20

Table 1. k=300 p-value=0.10

Tests	$mle\lambda_0$	$95\%mle\lambda_0$	$k(1-mle\lambda_0)$	PTP	PFP	PFN	EDR
PEV	0.53	(.46,.60)	141	0.90	0.10	0.002	0.998
PUV	0.53	(.47, .59)	141	0.90	0.10	0.001	0.998
BEV	0.47	(.38, .56)	159	0.90	0.10	0.002	0.990
BUV	0.47	(.38, .56)	159	0.90	0.10	0.002	0.990

4. CONCULSION

1. All p-value frequency plots indicate a non uniform distribution across p-values. Formal Q-test not shown here.

2. Assuming a single beta there are 140 to 159 genes for which there is a true difference in gene expression. The permutation tests give a more conservative estimate than the bootstrap.

3. For p-value of interest of 0.10, the PTP, PFP, PFN and EDR are consistent across all tests. All these values are in optimal ranges for the information sought in this data.

4. REFERENCES

Allison,D.B., Gadbury, G.L., Moonseong, H. Et Al. (2002) A Mixture Model Approach For The Analysis Of Microarray Gene Expression Data. Computational Statistics and Data Analysis. 39. Pp 1-20.

Brand, , J.P., Beasley, T.M., Bartolucci, A.A. Et Al. (2006) A Comparison Of Parametric And Non-Parametric Tests For Inferential Testing In Microarray Research. Paper Submitted.

Parker, R. A. And Rothenberg, R.B. (1988) Identifying Important Results From Multiple Statistical Tests. Statistics In Medicine. 7. Pp 1031-1043.