CHAPTER 6

Advanced Statistical Techniques for
Multiple Comparisons

6.1 Introduction

In recent years multiple testing has emerged as a very important topic in statistical inference, because of its applicability in understanding large data sets. One of the major fields of applications where multiple testing is extensively applied is the analysis of microarray gene expression. An important and common problem in microarray experiments is the detection of genes that are differentially expressed in a given number of classes. As this problem concerns the selection of significant genes from a large pool of candidate genes, it needs to be carried out within the framework of multiple hypothesis testing. Due to the cost, it is common that thousands of genes are measured with a small number of replications, as a con-

Some results included in this chapter is from Bindu et al. (2012(e)).
sequence, one faces a large $G$, small $n$ problem, where $G$ is the total number of genes and $n$ is the number of replications. In this chapter, we develop two methods to handle the multiplicity issue. The first one is based on the generalized $p$-value approach. With this approach, a measure of the FDR is provided for each gene. The second one is based on mixed Laplace model for Empirical Bayes inference. An attractive feature of the mixture model approach is that it provides a framework for the estimation of the prior probability that a gene is not differentially expressed, and this probability can subsequently be used in forming a decision rule. The rule can also be formed to take the false negative rate into account.

6.2 Multiple Testing in Microarray Using Generalized $P$-value

The biological question of differential expression can be considered as a problem in multiple hypothesis testing in which $m$ null hypotheses were simultaneously tested, where $m$ (the number of genes whose expression levels were measured) can be considerably large. The large number of genes on a microarray, will lead to the identification of many genes that truly are not differentially expressed (false discoveries). In such situations, false discoveries (true null hypothesis declared significant) are inevitable. Thus, it is important in any multiple testing problem to control the error rate of false discoveries. Multiple testing procedures consist of choosing a vector of cutoffs for the test statistics such that a suitably defined false positive rate is controlled at an a priori specified level $\alpha$. A standard approach to the multiple testing problem consists of two aspects:

1. Computing a test statistic $T_j$ for each gene $j$.

2. Applying multiple testing procedures to determine which hypotheses to reject while controlling a suitably defined Type I error rate.

Let $H_1, \ldots, H_m$ be $m$ independent hypotheses to be tested. Let $R_T$ be the number of true hypotheses that are incorrectly rejected, and let $R_N$ be the number of not true hypotheses that are rejected. The total number of hypotheses rejected is $R = R_T + R_N$. Assume that the $m_0$ out of the $m$ null hypothesis is true ($m_0$ genes are
Table 6.1: Type I and Type II errors in multiple hypothesis testing.

<table>
<thead>
<tr>
<th>Not rejected</th>
<th>Rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>$m_0 - R_T$</td>
<td>$R_T$</td>
</tr>
<tr>
<td>$H_0$ false</td>
<td>$m_1 - R_N$</td>
<td>$R_N$</td>
</tr>
<tr>
<td>Total</td>
<td>$m - R$</td>
<td>$R$</td>
</tr>
</tbody>
</table>

not differentially expressed) and $m_1 = m - m_0$ null hypothesis are false. A statistical test is performed independently, let $p_i$, $i = 1, 2, \ldots, m$ be corresponding p–values. The decision to reject (or not) can be correct or false; when the null hypothesis is rejected for one of the $m_0$ variables for which it is actually true, is the false discovery (type I error). Table 1 presents the possible categories to which rejected and non-rejected hypotheses can belong, and the number of hypotheses in each category.

A fundamental issue for multiple hypothesis testing is how to effectively control Type I errors, namely the errors of rejecting null hypotheses that are actually true. The problem of simultaneously testing a large number of hypotheses has generated a great amount of interest. Benjamini and Hochberg (1995) introduced the concept of False Discovery Rate (FDR). False discovery rate (FDR), is the expected proportion of Type I errors among the rejected hypotheses is given by,

$$\text{FDR} = E\left(\frac{R_T}{R} ; R > 0\right) = E\left(\frac{R_T}{R} | R > 0\right) \times Pr(R > 0).$$

In the present study we use the generalized p–value (Tsui and Weerahadi (1985)) approach for dealing the multiple hypothesis testing in yeast *Saccharomyces cerevisiae* microarray responding to diverse environmental transitions. The procedure for controlling the FDR in testing several independent null hypotheses based on generalized p-values is given as follows.
Let $H_1, \ldots, H_m$ are independent null hypotheses to be tested and $p_1, \ldots, p_m$ be the corresponding generalized p–values. Let $q$ be the FDR, then for a given $q > 0$, suppose that there exists a cumulative distribution function $F_i$ such that:

$$P(p_i \leq r | H_i) \leq F_i(r), \text{ for } r \leq r_0, \quad (6.2.1)$$

for some $r_0$ satisfying $q \leq F_i(r_0)$, assuming that $H_i$ is a true hypothesis. Let $p_i^* = F_i(p_i)$, for $i = 1, \ldots, m$ and $p_{(1)^*}^* \leq \ldots \leq p_{(m)^*}^*$ be the ordered values of $p_i^*$. Let $H_{(1)}, \ldots, H_{(m)}$ be the corresponding hypotheses. Define $q_i = \frac{iq}{m}$ and

$$k^* = \max\{i : p_{(i)}^* \leq q_i\}. \quad (6.2.2)$$

Then, the procedure that rejects $H_{(i)}$ for $i \leq k^*$ guarantees that $FDR \leq q$.

### 6.2.1 Generalized p-value for Inverse Gaussian

The microarray non log transformed data can be modeled using lognormal, gamma, Inverse Gamma and Inverse Gaussian ($IG$) distributions. In the present study we assume the Inverse gaussian distribution for non-log transformed data. The advantages of $IG$ model are: (1) they can accommodate a variety of shapes, from highly skewed to almost normal; (2) they are unique among the distributions for positively right-skewed data (e.g. Weibull, gamma, lognormal) due to the fact that they share many elegant and convenient properties with Gaussian models.

Krishnamoorthi and Tian (2008) developed generalized inferences on the difference and ratio of the means of two inverse Gaussian distributions. We developed the multiple hypothesis testing procedure based the generalized p-value approach under the assumption of $IG$ distribution. Then differential expression analysis of microarray gene expression data involves the comparison of two Inverse Gaussian means which can be evaluated by the generalized p–value method. Comparison of the means of transformed data in two samples can produce a different conclusion as opposed to comparing the means of the original data. The
large sample test is too liberal where as the test based on the generalized p–
values controls the type–I error quite satisfactorily. The procedure to be applied on
untransformed data is summarized as follows:

Let $X_{gij}, i = 1, 2; j = 1, 2, ..., n_i$ and $g = 1, 2, ..., n_g$ denote the random samples
of gene expression data. Let $X'_{gij}$ follow two parameter Inverse Gaussian($IG$)
distributions with parameters $\mu_i$ and $\sigma_i$, for $i = 1, 2$. Then the problem of testing
the differential expression reduce to the testing of the equality means of two $IG$
distribution.

Let $X$ follow a two parameter Inverse Gaussian family, then the
pdf of $X$ is given
by
\[
f(x; \mu, \sigma) = \left( \frac{\sigma}{2\pi x^3} \right)^{1/2} \exp \left( -\frac{\sigma (X - \mu)^2}{2\mu^2 x} \right), X > 0, \mu > 0, \sigma > 0, \tag{6.2.3}\]
where $\mu$ is the location parameter and $\sigma$ is the scale parameter.

Let $X_1, ..., X_n$ be a sample of observations from an $IG$ distribution with pdf
in (6.2.3). Let $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ and $V = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{X_i}{\bar{X}} - \frac{1}{\bar{X}} \right)$. $\bar{X}$ and $V$ are mutually
independent and $\bar{X} \sim IG(\mu, n\sigma)$ and $n\sigma V \sim \chi^2_{n-1}$. Then $\bar{X} \sim IG(\mu, n\sigma)$ and $n\sigma V \sim \chi^2_{n-1}$.

Then the generalize pivotal variable for $\mu$ is given by
\[
T_\mu = \frac{\bar{x}}{\max \left( 0, 1 + \sqrt{\frac{(n-1)(X/\mu-1)}{VX}} \sqrt{\frac{\sigma^2}{n-1}} \right)} = \frac{\bar{x}}{\max \left( 0, 1 + W \sqrt{\frac{\sigma^2}{n-1}} \right)}, \tag{6.2.4}
\]
where $W = \frac{\sqrt{(n-1)(X/\mu-1)}}{VX}$ and $|W| \sim |t_{n-1}|$.

Let $X$ be an $IG$ variable with pdf $f(x; \mu_1, \sigma_1)$ and $Y$ be an $IG$ variable with pdf
$f(y; \mu_2, \sigma_2)$, with pdf given in (6.2.3). Let $X_1, ..., X_{n_1}$ be a sample of observations
from $X$ and $Y_1, ..., Y_{n_2}$ be a sample of observations from $Y$. The the generalized
pivotal variable for the difference of means of $IG$ distributions is
\[
T^*_g = T_{\mu_1} - T_{\mu_2}, \tag{6.2.5}
\]
where $T_{\mu_1}$ is obtained by putting in $\mu = \mu_1$ in equation (6.2.4). Now consider the problem of testing

$$H_0: \mu_1 = \mu_2 \text{ vs } \mu_1 \neq \mu_2.$$  

(6.2.6)

The generalized test variable for testing $H_0: \mu_1 = \mu_2$ is $T_g = T_g^* - (\mu_1 - \mu_2)$, where $T_g^* = T_{\mu_1} - T_{\mu_2}$. The hypothesis will be rejected for a small p-value for comparison of either of the two alternative hypothesis. Hence the generalized p-value for the two sided test can be obtained as:

$$2 \times \min \{P(T_g \leq 0), P(T_g \geq 0)\}.$$  

(6.2.7)

### 6.2.2 Estimation of FDR

Let $N$ be the total number of genes, $G$ be the total number of significant genes and $p_i$ is the generalized p-value of the $i^{th}$ most significant gene as estimated from generalized p-value technique. Then the FDR can be estimated using the following formula.

$$FDR_G = \min_{i \geq G} \left[ \frac{N \times p_i}{i} \right].$$  

(6.2.8)

### 6.2.3 Yeast Microarray Data Analysis

We have downloaded the complete data set from http://www-genome.stanford.edu/yeast stress. In the present work we explored the data on genomic expression patterns in the yeast *Saccharomyces cerevisiae* responding to heat shock out of the different environmental parameters on the expression levels of 6200 genes studied by Gasch, et al. (2000). In the present study the multiple testing problem based on the generalized p–value is developed to identify the differentially expressed genes.

### 6.2.4 Results and Discussion

In the present work we explored the data on genomic expression patterns in the yeast *Saccharomyces cerevisiae* responding to heat shock. Cells subjected to a larger shift in temperature responded with larger and more prolonged alterations in gene expression before adapting to their new steady-state expression levels, relative
to cells exposed to smaller temperature changes. One group consisted of genes whose transcript levels increased in abundance in response to the environmental changes, and the other group was comprised of genes whose transcript levels decreased following environmental stress. The genes whose transcript levels increase in response to environmental change will be referred to as induced, while genes whose transcript levels decrease will be referred to as repressed. A large set of genes ($\sim 900$) showed a similar drastic response to these environmental changes.

We have performed testing for differentially expressed genes using the generalized method for the downloaded datasets from Gasch, et al. (2000). After getting the raw p-values for individual gene by using either the t-test (unequal variance) or by the generalized p-value method, p-values were adjusted for the multiplicity problem by fixing different q values using the equation (6.2.2). The estimated false discovery rate were computed and are given in Table 2. The generalized p-value method could find more number of genes at a particular level of q from the repressed dataset as well as in induced dataset. By reducing the level of q we can go for stringent selection of genes but will increase the chance of not detecting truly expressed genes. In general the generalized p value method could identify more number of truly expressed genes with a fewer chance of accepting false discoveries. We computed expected false discovery rate by selecting 200 to 900 genes from the control dataset and test dataset.

In this study the concept of generalized p-value method has been applied for the selection of differentially expressed genes in heat shock treatment. We identified 900 genes altered in response to heat shock with less false discovery rate. The results of generalized p-value is compared with the results of t-test approach by assuming unequal sample variances. The numerical results show that the procedure based on generalized p-value is superior to the t-test. Similar approach may be used to identify differentially expressed genes in other treatment conditions described in the Gasch, et al. (2000) or other microarray experiments. The
**Table 6.2:** The estimated false discovery rates based on t–test and generalized p-value technique

findings show that this statistical modeling approach based on generalized p–value provides an improved method for gene selection with low number of false positives.

The analysis of microarray using multiple testing by controlling the FDR may reduce the power available to detect changes in expression for individual genes (Ge et al. (2002)). On the other hand, the parallel nature of the inference in microarrays allows some compensating possibilities for borrowing information from the ensemble of genes which can assist in inference about each gene individually. One way do this is through the applications of Empirical Bayes methods (Efron et al. (2001 and 2003)).

### 6.3 Empirical Bayes Approach for Analysing Differentially Expressed Genes in Microarray

The Empirical Bayes approach is very popular in the context of analyzing the microarray data. It is well understood that due to a very small number of replicates standard maximum likelihood estimates of the variance of the individual gene ex-
pression are very imprecise. Therefore the Bayes hierarchical model is often used and the standard deviation for each of the respective t-statistics is estimated based on the Empirical Bayes approach. The method is implemented in the package LIMMA (Smyth (2004)), which became a standard tool for practical microarray analysis. Empirical Bayes methods replaces the hyper-parameters of a Bayes procedure by maximum likelihood, method of moments or other estimates from the data. These methods allow one to estimate quantities (probabilities, averages, etc.) about an individual member of a population by combining information from the individual and other subjects in the empirical study.

6.3.1 Two Component Mixed Laplace Distribution

Many contemporary multiple testing procedures are based on a two component mixture model (Efron et al. (2001), Efron and Tibshirani (2002) and McLachlan et al. (2006)). Such a mixture model assumes that test statistics $X_1, \cdots, X_m$ are independent and identically distributed and their marginal $cdf$ $F(x)$ can be given by,

$$ F(x) = p_0 F_0(x) + p_1 F_1(x), \quad (6.3.1) $$

where $F_0(x)$ and $F_1(x)$ denote the cdfs of the null and alternative, respectively. The $p_0$ and $p_1$ are respectively, the expected proportion of null and alternatives among all tests. In the present Chapter we discuss multiple hypothesis testing using the two component Laplace mixtures.

A two-component Laplace mixture model can be defined by two Laplace distributions, $L(\mu_1, \sigma_1)$ and $L(\mu_2, \sigma_2)$, and the probability that the random variable (the observable) arises from the first distribution is $\kappa$. The parameter in this model is the vector $\theta = (\kappa, \mu_1, \sigma_1, \mu_2, \sigma_2)$. Then the pdf of the mixture is

$$ f(x; \theta) = \kappa f_1(\mu_1, \sigma_1) + (1 - \kappa) f_2(\mu_2, \sigma_2), $$

where $f_i(\mu_i, \sigma_i)$ is the Laplace pdf with parameters $\mu_i$ and $\sigma_i$. 

128
Chapter 6. Advanced Statistical Techniques for Multiple Comparisons

Figure 6.1: The plot of $MLD(\mu_1, \mu_2, \sigma_1, \sigma_2, \kappa)$ for $\mu_1 = 2$, $\mu_2 = -2$, $\sigma_1 = \sigma_2 = 1$ and for various values of $\kappa$.

The mixture of distributions adaptively gives flexible shapes of density functions over multi-modality and non-normality. The density plot of mixed Laplace distribution for various values of $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \kappa)$ are shown in Figure 6.1, Figure 6.2 and Figure 6.3.

Figure 6.1, illustrate the behaviour of MLD for various values of $\kappa$. Figure 6.2 and Figure 6.3 depicts the effect of the scale parameters and the location parameters respectively. From these figures we can see that mixed Laplace density is capable of accounting multi-modality, skewness and heavy tail characteristics in the data.

6.3.2 Likelihood and Estimation of $MLD$

Let $X = (X_1, X_2, \ldots, X_n)$ be independent and identically distributed samples from the two component mixed Laplace distribution. Then the log-likelihood function for
**Figure 6.2:** The plot of $MLD(\mu_1, \mu_2, \sigma_1, \sigma_2, \kappa)$ for $\mu_1 = 2$, $\mu_2 = -2$, $\kappa = 0.5$ and for various values of $\sigma_1$ and $\sigma_2$.

**Figure 6.3:** The plot of $MLD(\mu_1, \mu_2, \sigma_1, \sigma_2, \kappa)$ for $\sigma_1 = 1$, $\sigma_2 = 0.5$, $\kappa = 0.5$ and for various values of $\mu_1$ and $\mu_2$. 
\( \theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \kappa) \) is of the form \( L(\theta) = \sum_{i=1}^{n} L_i(\theta) \) where

\[
L_i(\theta) = \log \left( \frac{\kappa}{2\sigma_1} e^{-\frac{|X_i - \mu_1|}{\sigma_1}} + \frac{1 - \kappa}{2\sigma_2} e^{-\frac{|X_i - \mu_2|}{\sigma_2}} \right).
\]

Maximum likelihood estimators (MLE) of the parameters \((\mu_1, \mu_2, \sigma_1, \sigma_2, \kappa)\) are obtained by solving four score equations, e.g., \( \frac{\partial \log L}{\partial \theta}(\mu_1, \mu_2, \sigma_1, \sigma_2, \kappa; X) = 0 \) and so on. Numerical methods are needed to solve these score equations. In our illustration, the maximization of the likelihood is implemented using the \textit{optim} function of the R statistical software, applying the BFGS algorithm (See R Development Core Team (2006)). Estimates of the standard errors were obtained by inverting the numerically differentiated information matrix at the maximum likelihood point.

### 6.3.3 Application to Microarray Data

In this section we apply the two component mixed Laplace distribution to model the distribution of a cDNA dual dye microarray gene expression data sets. We downloaded two cDNA dual dye microarray data sets (Experiment id-51398 and id-51401) from the Stanford Microarray Database. Each array chip contains approximately 42,000 human cDNA elements, representing over 30,000 unique genes. The dataset was normalized using LOWESS normalization method.

We fitted the mixed Laplace distribution (MLD) and mixed Gaussian distribution (MGD) to the intensities \( m = \log_2(R_i/G_i) \) for the data sets (Experiment id-51398 and Experiment id-51401). We obtained maximum likelihood estimates and their asymptotic standard errors for the parameters of \( MLD(\mu_1, \mu_2, \sigma_1, \sigma_2, \kappa) \) and \( MGD(\mu_1, \mu_2, \sigma_1, \sigma_2, \kappa) \) and are reported in Table 6.3. Figure 6.3 depicts the histograms of the gene expression data sets and the fitted probability density function evaluated at the MLEs.
Figure 6.4: Fitted mixed Laplace density (red line) and mixed Gaussian density (black line) evaluated at MLEs superimposed on histogram of the microarray data, left panel: Experiment id-51398, right panel: Experiment id-51401.

Table 6.3: Microarray data analyses - Maximum likelihood estimates and their asymptotical standard deviations for mixed Laplace (MLD) and mixed Gaussian distributions (MGD) for two datasets.

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\mu}_1$</th>
<th>$\hat{\mu}_2$</th>
<th>$\hat{\sigma}_1$</th>
<th>$\hat{\sigma}_2$</th>
<th>$\hat{\kappa}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLD</td>
<td>id-51398: -0.11 (0.007)</td>
<td>id-51401: 0.03 (0.011)</td>
<td>id-51398: 0.70 (0.010)</td>
<td>id-51401: 1.22 (0.003)</td>
<td>id-51398: 0.52 (0.013)</td>
</tr>
<tr>
<td></td>
<td>id-51398: -0.04 (0.002)</td>
<td>id-51401: -0.01 (0.002)</td>
<td>id-51398: 0.20 (0.005)</td>
<td>id-51401: 0.27 (0.003)</td>
<td>id-51398: 0.35 (0.005)</td>
</tr>
<tr>
<td>MGD</td>
<td>id-51398: -0.02 (0.010)</td>
<td>id-51401: 0.11 (0.004)</td>
<td>id-51398: 1.24 (0.010)</td>
<td>id-51401: 0.18 (0.006)</td>
<td>id-51398: 0.38 (0.006)</td>
</tr>
<tr>
<td></td>
<td>id-51398: 0.02 (0.003)</td>
<td>id-51401: -0.11 (0.008)</td>
<td>id-51398: 0.29 (0.003)</td>
<td>id-51401: 0.67 (0.009)</td>
<td>id-51398: 0.37 (0.014)</td>
</tr>
</tbody>
</table>

132
### Table 6.4: Microarray data analyses - Akaike’s Information Criterion (AIC) and Bayesian Information Criterion (BIC) for assessing goodness-of-fit of mixed Laplace density (MLD), and mixed Gaussian density (MGD).

<table>
<thead>
<tr>
<th></th>
<th>51398</th>
<th>51401</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AIC</td>
<td>BIC</td>
</tr>
<tr>
<td>MLD</td>
<td>74390.5</td>
<td>74433.7</td>
</tr>
<tr>
<td>MGD</td>
<td>75622.4</td>
<td>75665.5</td>
</tr>
</tbody>
</table>

Anderson (1998)) and Bayesian Information Criterion (BIC) (Schwarz (1978)) to assess the appropriateness of MLD over the MGD. The $AIC$ and $BIC$ are given by

\[
AIC = -2 \log L + 2K \quad \text{and} \quad BIC = -2 \log L + K \log(n),
\]

where $\log L = \log(L_f(\hat{\theta}|x_1, \ldots, x_n))$ is the log-likelihood of the data $x_1, \ldots, x_n$ under the probability distribution $f$, $K$ is the number of parameters being estimated, $\hat{\theta}$ is the maximum likelihood estimate of the parameters of $f$ and $n$ is the sample size. Note that $AIC$ does not explicitly take into account the sample size as $BIC$ does. However, in most cases $AIC$ and $BIC$ are of similar nature and give consistent results for model selection. A smaller value of $AIC$ or $BIC$ indicates a better fit. Table 6.4 shows the $AIC$ and $BIC$ for the $MLD(\mu_1, \mu_2, \sigma_1, \sigma_2, \kappa)$, and $MGD(\mu_1, \mu_2, \sigma_1, \sigma_2, \kappa)$ distributions for the two datasets examined. The $MLD(\mu_1, \mu_2, \sigma_1, \sigma_2, \kappa)$ distribution had a lower $AIC$ and $BIC$ for two datasets. A smaller value of $AIC$ and $BIC$ indicates a better fit, and hence, $MLD$ seems to fit the data better than $MGD$. From Table 6.4 we can see that $AIC_{MLD} - AIC_{MGD} < 0$ and $BIC_{MLD} - BIC_{MGD} < 0$.

We have proposed the mixed Laplace distribution as a class of distributions that might be used to describe the pattern of differential gene expression for genes that are differentially expressed. This family was chosen primarily for its flexibility and tractability.
6.3.4 Multiple Testing of Microarray via Laplace Mixture Model

In microarray experiments there are often relatively few arrays, resulting in highly variable estimates of the standard deviation for each gene. To use the information in the large number of measured genes to handle this problem, an empirical Bayes approach (Maritz and Lwin (1989), Baldi and Long (2001), Lönnstedt and Speed (2002) and in the R (R Development Core Team, 2006) package LIMMA (Smyth, 2004)) can be used.

After preprocessing of the raw image data, the expression levels are often assumed to follow a two-groups model, that is, the expressions are each either null or nonnull with prior probability $\pi_0$ or $\pi_1 = 1 - \pi_0$, respectively. The two-groups model plays an important role in the Bayesian microarray literature and is broadly applicable (Efron, 2008). McLachlan et al. (2006) developed two-component Gaussian mixture framework, using an empirical Bayes approach for the detection of genes that are differentially expressed. In this chapter we discuss some preliminary aspects of two-component mixed Laplace framework using empirical Bayes approach for detecting the differentially expressed genes.

6.3.5 Bayesian Framework of Laplace Mixed Model

In this section we develop a decision-theoretic approach to the problem of finding genes that are differentially expressed by following the approach of McLachlan et al. (2006). We developed a prediction rule approach based on a two component Laplace mixture model,

$$f(x; \theta) = \pi_0 f_0(\mu_0, \sigma) + \pi_1 f_1(\mu_1, \sigma).$$

Where $\pi_0$ and $\pi_1$ are the prior probabilities to belong to groups not differentially expressed and differentially expressed, respectively. We assume that the density $f_i$ is Laplace with mean $\mu_i$ and standard deviation $\sigma$ (variances are assumed to be homogeneous). Here, the variances ($\sigma^2$) of the genes are assumed to be inverse gamma distributed. The two hyper parameters $\alpha$ and $\beta$ of the inverse gamma
distribution are then estimated by maximizing the marginal likelihood using the nlminb function in R. The other unknown parameters are can be estimated using the EM algorithm.

Let \( g \) denote the population of genes under consideration. It can be decomposed into \( g_0 \) and \( g_1 \), where \( g_0 \) is the population of genes that are not differentially expressed, and \( g_1 \) contains the genes that are differentially expressed. We let the random variable \( X_{ij} \) be defined to be one or zero according as the \( j^{th} \) gene belongs to \( g_i \) or not \((i = 0, 1; j = 1, \ldots, N)\). We define \( H_j \) to be zero or one according as to whether the null hypothesis of no differential expression does or does not hold for the \( j^{th} \) gene. Thus \( X_{1j} \) is zero or one according as to whether \( H_j \) is zero or one.

The prior probability that the \( j^{th} \) gene belongs to \( g_0 \) is assumed to be \( \pi_0 \) for all \( j \). That is, \( \pi_0 = Pr(H_j = 0) \) and \( \pi_1 = Pr(H_j = 1) \). Assuming that the test statistics \( W_j \) (posterior t-statistics) have the same distribution in \( g_i \). Let \( f_i(w_j) \) denote the density of \( W_j \) in \( g_i \) \((i = 1, 2)\). The unconditional density \( f(w_j) \) of \( W_j \) is given by the two-component Laplace mixture model.

\[
f(w_j) = \pi_0 f_0(w_j) + \pi_1 f_1(w_j).
\] (6.3.2)

Using Bayes Theorem, the posterior probability that the \( j^{th} \) gene is not differentially expressed (that is, belongs to \( g_0 \)) is given by

\[
\tau_0(w_j) = \frac{\pi_0 f_0(w_j)}{f(w_j)} (j = 1, \ldots, N).
\] (6.3.3)

The posterior probability that the \( j^{th} \) gene is differentially expressed (that is, belongs to \( g_1 \)) is given by

\[
\tau_1(w_j) = 1 - \frac{\pi_0 f_0(w_j)}{f(w_j)} (j = 1, \ldots, N).
\] (6.3.4)

In this framework, the gene-specific posterior probabilities \( \tau_0(w_j) \) provide the basis
for optimal statistical inference about differential expression.

Let $e_{01}$ and $e_{10}$ denote the two errors when a rule is used to assign a gene to either $g_0$ or $g_1$, where $e_{ij}$ is the probability that a gene from $g_i$ is assigned to $g_j$ ($i, j = 0, 1$). That is, $e_{01}$ is the probability of a false positive and $e_{10}$ is the probability of a false negative. Then the risk is given by

$$Risk = (1 - c)\pi_0 e_{01} + c\pi_1 e_{10},$$

(6.3.5)

where $(1 - c)$ is the cost of a false positive. As the risk depends only on the ratio of the costs of miss allocation, they have been scaled to add to one without loss of generality. The Bayes rule, which is the rule that minimizes the risk (6.3.4), assigns a gene to $g_1$ if

$$\tau_0(w_j) = c;$$

(6.3.6)

otherwise, the $j^{th}$ gene is assigned to $g_0$. In the case of equal costs of miss allocation ($c = 0.5$), the cutoff point for the posterior probability $\tau_0(w_j)$ in (6.3.5) reduces to 0.5.

**Estimation of FDR**

If $\hat{\pi}_0$, $\hat{f}_0(w_j)$, and $\hat{f}(w_j)$ denote estimates of $\pi_0$, $f_0(w_j)$, and $f(w_j)$, respectively, the gene-specific summaries of differential expression can be expressed in terms of the estimated posterior probabilities $\hat{\tau}_0(w_j)$, where

$$\hat{\tau}_0(w_j) = \frac{\hat{\pi}_0 \hat{f}_0(w_j)}{\hat{f}(w_j)} (j = 1, \ldots, N).$$

(6.3.7)

is the estimated posterior probability that the $j^{th}$ gene is not differentially expressed. An optimal ranking of the genes can therefore be obtained by ranking the genes according to the $\hat{\tau}_0(w_j)$ ranked from smallest to largest. A short list of genes can be obtained by including all genes with $\hat{\tau}_0(w_j)$ less than some threshold $c_o$ or by taking
the top No. of genes in the ranked list. Suppose that we select all genes with
\[ \hat{\tau}_0(w_j) \leq c_0. \] (6.3.8)

Then an estimate of the local FDR rate is given by
\[ \hat{FDR} = \frac{\sum_{j=1}^{N} \hat{\tau}_0(w_j)I_{[0,c_0]}(\hat{\tau}_0(w_j))}{N_r}, \] (6.3.9)

where \( N_r = \sum_{j=1}^{N} \hat{\tau}_0(w_j)I_{[0,c_0]}(\hat{\tau}_0(w_j)) \) is the number of the selected genes in the list. Here \( I_A(w) \) is the indicator function that is one if \( w \) belongs to the interval \( A \) and is zero otherwise. Thus we can find a data-dependent \( c_0 \leq 1 \) as large as possible such that \( \hat{FDR} = a \). This assumes that there will be some genes with \( \hat{\tau}_0(w_j) \leq a \), which will be true in the typical situation in practice.

To check the performance of the two component mixed-Laplace model we used the colon cancer microarray datasets. We used the R Package for the analysis.

### 6.3.6 Application to Colon Cancer Data

We have downloaded the Colon Cancer Data of Alon et al. (1999). Alon et al. (1999) have presented a data set that contains gene expression levels of 40 tumour and 22 normal colon tissues for 6500 human genes obtained with an Affymetrix oligonucleotide array. We assumed the (pooled) t-statistic as our test statistic \( W_j \) for each gene. We analyzed the dataset and estimated the probability of differential expression. Using empirical Bayes methods and assuming the two-component Laplace mixture model, we obtained \( \pi_0 = 0.35, \mu_0 = -0.01, \mu_1 = 0.15, \alpha = 9.5 \) and \( \beta = 0.15 \). The near-zero mean of the nonnull mixture component suggests that there may be two nonnull groups (over- and under-expressed groups of genes).

We fitted the Laplace mixed model to the data. Using null posterior probability threshold of 0.2 we detect 175 nonnull genes, while using the FDR method (with a threshold of 0.2) we get 150 genes. Detecting nonnull genes in a typical microarray gene expression analysis involves setting a minimum fold-change threshold, in
addition to setting the level at which the False Discovery Rate is controlled.

This preliminary study needs to be further developed. Also a comparative study with the existing study is worth investigating and studies in this direction are ongoing.

### 6.4 Conclusion

In this paper we presented two approaches for performing multiple hypothesis testing in microarray. We developed a generalized p-value approach by controlling the false discovery rate and a unified modelling framework for empirical Bayes inference in microarray experiments. Our model-based approach also allows one to declare the nonnull status of a gene by controlling the false discovery rate (FDR). The Empirical Bayes procedure described in this Chapter provides an effective framework for studying the relative changes in gene expression for a large number of genes.

The promise of gene expression studies using microarray technology has inspired much new hope for finding complex disease genes. The majority of the initial technical challenges of conducting experiments are being resolved. Progress in the genomics revolution is limited by our ability to transform such large amounts of raw data into reliable and meaningful biological sense. Improvement in statistical tools in data analysis is necessary to draw valid information out of the data. Advanced software and statistical tools addresses the multidimensional complexities faced by investigators in drawing meaningful conclusions from their microarray data.

**References**


Alon, Barkai, Noterman, Gish, Ybarra, Mack and Levine (1999). Broad pat-


