Quantitative methods to measure gene expression levels have been used by biologists for more than twenty years. Northern and southern blots are techniques used to identify and locate mRNA and DNA sequences that are complementary to a segment of DNA (Alwin et al., (1977); White, (1995)). Serial Analysis of Gene Expression (SAGE) is able to measure the global gene expression from entire cell or tissue (Velculescu et al., (1995)). Microarray technology provides biologists the ability to measure the expression levels of thousands of genes simultaneously (Lander, 1999). Two key concepts behind this measurement process are reverse transcription and hybridization. Currently two platforms have emerged to dominate the microarray field, oligonucleotide and spotted cDNA arrays. The first approach developed by the Affymatrix group involves the in situ synthesis of oligonucleotides (less than 30 bp long) onto solid substrates using photo lithographic techniques (Lipshutz et al., (1999)). The second platform developed at Stanford University
involves robotically printing cDNA clone inserts (200 to 2000 bp long) onto glass microscope slides Brown and Botstein (1999). The supplement to Nature Genetics, *The Chipping Forecast* and the books “DNA Microarrays : A Practical Approach”, “Microarray Biochip Technology” and “Microarrays- Chipping in Genomics” provide general overviews of microarray technologies and different areas of application of microarrays (Schena (1999, 2000)). Image data from the arrays leads to gene specific numerical intensities representing the relative expression levels and these in turn form the input to computational analysis designed to assess the significance and relationship across biological conditions. The promise of this technology is the ability to observe the entire gene in action and being doing so, to uncover its underlying expression mechanism. Some applications of this technology are: deducting functions of unknown genes, identifying disease profiles, deciphering regulatory mechanisms, classification of biological conditions, genotyping and drug development.

Novel computation tools and reliable data processing procedures are essential for the meaningful and accurate interpretation of microarray data (Quackenbush (2001)). The challenge faced by bioinformatician in analysis is that the data contains overwhelming number of genes compared to the number of samples. Analysis of these data requires statistical tools that is acceptable to high dimensional data to make them more comprehensive. The common strategy to reduce data variability and data dimensionality is to perform two preprocessing operations before undertaking any analysis of the data: normalization and filtering. Normalization removes systematic distortions across microarrays and filtering reduce variability and decrease the dimensionality of the data by removing genes that are not sufficiently differentiated.

Fold change is the simplest method for identifying differentially expressed genes (De Risi et al. (1997 ); Eisen et al., (1998)). It is based on the observed ratio (or
average of ratios) between two conditions. An arbitrary cut-off value (for example, 2 fold) is used to identify differentially expressed genes. Fold change gives no assessment of statistical significance. Identifying differentially expressed genes involves the selection of test statistic which will rank the genes in order of strength of evidence of differential expression and to choose a critical value for ranking of statistic above which any value is considered to be significant. The standard $t$–test is usually used to identify significantly differentiated genes across two conditions, which assumes normally distributed data and equal variance within classes. Welch statistics for unequal variances, Wilcoxon nonparametric test, permutation test, penalized $t$-statistics (Efron et al., (2001)), significance analysis of microarrays (SAM) (Tusher et al., (2001)), regularized $t$-test (Cyber-T) (Baldi and Long (2001)), moderated $t$-statistics (Smyth (2004)) are also applied for identifying differentially expressed genes in two conditions. Cui and Churchill, (2003) have given a review of test statistics for differential expression for microarray experiments. Under multiple conditions and different sources of variation ANOVA method and mixed model method are applied (Kerr et al., (2000); Wolfinger et al., (2001)). Other approaches to linear models for microarray data analysis have been described by Chu et al., (2002), Yang and Speed, (2002) and Yang and Speed, (2003).

2.1 Representation of gene expression data

Gene expression data can be represented as a real matrix, each row in the matrix contains the expression data regarding a specified gene, and each column represents a condition, different time periods or tissue profile. Let the matrix $X$, of size $(m \times n, m$-genes $\times n$-arrays), tabulate the full expression data, where $x_{ij}$ is the log$_2$ of the expression ratio of the $i^{th}$ gene in the $j^{th}$ sample as measured by the array $j$. The vector in the $i^{th}$ row of the matrix $X$ lists the fluorescence ratio of the $i^{th}$ gene across different samples and are called the gene expression profile for $i^{th}$ gene. The
vector in the $j^{th}$ column of the matrix $X$, lists the genome wide fluorescence ratios measured by the $j^{th}$ array.

Alternative approach with stronger assumption treats the data from single channel intensities Red (R) and Green (G) as primary data (*Kerr et al.*, (2000), *Wolfinger et al.*, (2001)).

### 2.2 Data preprocessing

A common strategy to reduce data variability and data dimensionality is to perform two preprocessing operations namely, *normalization* and *filtering* before undertaking any analysis of the data. The goal of the first operation is to remove systematic distortions across microarrays in order to render comparable experiments conducted under different conditions. The aim of the filtering operation is two fold: to reduce variability by removing those genes whose measurements are not sufficiently accurate and to decrease the dimensionality of the data by removing genes that are not sufficiently differentiated.

#### 2.2.1 Normalization of Microarray Data

One well known problem of cDNA technology is the consistent unbalance of the fluorescent intensities of the two dyes Cy3 (G) and Cy5 (R). There are several normalization methods and little consensus about which one to use. The choice of the method is coupled with the experimental design and these methods to reduce experimental errors can be carried out within each array and between arrays. *Kerr et al.*, (2000) and *Wolfinger et al.*, (2001) used ANOVA models for normalization where the “normalized” microarray data provide estimates of changes in gene expression that are corrected for potential confounding effects. *Chen et al.*, (1997) used the ratio based method of normalization. They assume that although individual genes might be up or down regulated, in closely related cells, the total quantity of RNA produced is approximately the same for essential genes. Using this assumption they
developed an approximate probability density ratio \( T_k = \frac{R_k}{G_k} \), where \( R_k \) and \( G_k \) are the measured red and green intensities for the \( k^{th} \) array. Another method suggested is to use a LOWESS (locally weighted scatterplot smoothing) approach for each print-tip group and its original application is for smoothing scatterplots in a weighted, least-squares fashion (Cleveland WS, (1979)). Because the LOWESS smoother is available in many statistical packages (R Development Core Team, 2005), the ideas of Dudoit et al., (2002) are easily implemented, and the two-step procedure for separately taking normalization and identification has been commonly accepted without question. Another approach is to build a regression curve to fit the intensity versus variance relationship (Jain, 2003).

### 2.2.2 Variance Stabilization

In a pioneering paper Chen et al., (1997) derived a model based on the assumption of a constant coefficient of variation, and derived the distribution of the ratio intensities. The distribution has one parameter, the coefficient of variation and according to the model is same for all the probes on the array. To fit their model to the intensity data from a two color cDNA array, they used a multiplicative calibration. These concepts have been widely used in microarray data analysis. However it has become clear that for many data sets that are encountered in practice they are insufficient and the limitation mostly affects the data from weakly expressed genes (Newton et al., (2001); Rocke and Durbin (2001); Rockey and Durbin (2003); Theilhaber et al., (2001)).

Huber et al., (2002) introduced a statistical model for microarray gene expression data that comprise data calibration, the quantification of differential expression and the quantification of measurement error. They derive a transformation \( h \) for intensity measurements and a difference statistic \( \Delta h \) whose variance is approximately constant along whole intensity range. For transformation \( h \), the parametric form \( h(x) = \arcsinh(a + bx) \) is derived from a model of the variance–versus–mean
dependant for the microarray intensity data, using the method of variance stabilizing transformations. For large intensities, \( h \) coincides with the log–ratio. Rocke and Durbin (2001, 2003) modelled the measured expression levels from microarray data as

\[
y = \alpha + \mu + e^n + \epsilon
\]

where, \( y \) is the measured raw expression level for a single color, \( \alpha \) is the mean background noise, \( \mu \) is the true expression level, and \( n \) and \( \epsilon \) are normally distributed error terms with mean 0 and variances \( \sigma_n^2 \) and \( \sigma_\epsilon^2 \) respectively. The variance of \( y \) under this model is,

\[
\text{Var}(y) = \mu^2 S_n^2 + \sigma_\epsilon^2
\]  

(2.1)

where \( S_n^2 = e^{\sigma_n^2 - 1} \). It is shown that for a random variable \( z \) satisfying \( \text{Var}(z) = a^2 + b^2 \mu^2 \), with \( \text{E}(y) = \mu \), there is a transformation that stabilizes the variance to the first order, meaning that the variance is almost constant no matter what the mean might be (Huber et al., 2002; Durbin et al., 2002)). There are several equivalent ways of writing this transformation for example,

\[
f_c(z) = \ln \left( \frac{z + z^2 + c^2}{2} \right)
\]

(2.2)

where \( c = \frac{a}{b} \). This transformation converges to \( \ln(z) \) for large \( z \), and is approximately linear at 0 Durbin et al., 2002). Since this is exactly the natural logarithm when \( c = 0 \), it was called the generalized logarithm or glog transformation. The inverse transformation is

\[
f_c^{-1}(w) = e^w - c^2 e^{-w} \]

(2.3)

Both \( f_c \) and its inverse are monotonic functions, defined for all values of \( z \) and \( w \), with derivatives of all orders. For array data, the random variable satisfies (exactly
or approximately)

\[ \text{Var}(z) = a^2 + b^2 \text{E}(z)^2 \]  

(2.4)

### 2.3 Visualization tools

There are various methods used for visualization of gene expression data.

#### 2.3.1 Box plots

A box plot is a plot that represents graphically several descriptive statistics of a given data sample (Fig 2.1). The method is usually used for finding outliers in the data. The box plot contains a central line and two tails. The central line in the box shows the position of the median. The box will represent an interval that contains 50 % of the data. Data points that fall beyond the box's boundaries are considered outliers. Such plots are often used in describing the range of log ratios that is associated with replicate spots.
2.3.2 Scatter plots

The scatter plot is a two or three dimensional plot in which a vector is plotted as a point having the coordinates equal to the components of the vector. Each axis corresponds to an experiment and each expression level corresponding to an individual gene is represented as a point. In such a plot, genes with similar expression levels will appear somewhere on the first diagonal (the line y=x) of the coordinate system. A gene that has an expression level that is very different between the two experiments will appear far from the diagonal. Therefore, it is easy to identify such genes very quickly. Scatter plots are easy to use but may require normalization of the data points in order to acquire accurate results. The most evident limitation of scatter plots is the fact that these can only be applied to data with two or three components since they can only be plotted in two or three dimensions. To overcome this problem the researcher may use the Principal Component Analysis (PCA) method.
2.3.3 Volcano plots

Volcano plot is a scatterplot showing the differential expression statistics and fold change. The name comes from the volcano shape of the plots. Volcano plots help to visualize effects of fold change and statistical significance simultaneously. Plots typically show log odds on the ordinate and fold-change values on the abscissa for all genes in a data set. The upper corners of the plot represent genes that show both statistical significance and large fold changes (Fig 2.2).

2.3.4 Heatmaps

Heatmaps offer a quick overview of clusters of genes that show similar expression values. Heat maps consist of small cells, each consisting of a color which represents relative expression values (Fig 2.3). Heatmaps are often generated from hierarchical cluster analyses of both samples and genes. Often the rows represent genes of similar expression values, whereas the columns indicate different biological samples.

2.3.5 MA Plots

MA plots are used to detect artifacts in the array that are intensity dependent. These are often used as an aid when normalizing two-colour cDNA microarrays. The data consist of intensity measurements that correspond to both red (R) and green (G) dyes. A Cartesian plot is constructed with M on the ordinate and A on the abscissa, where $M = \log_2 \frac{R}{G}$ and $A = \log_2 (\sqrt{RG})$. The data are often fitted with a lowess curve, which is used to normalize the gene-expression measurements (Fig 2.4).

2.4 Analysis of Comparative Experiments

2.4.1 Fold change

A simple microarray experiment is carried out to detect the differences in expression under the two conditions. Each condition may be represented by one or more RNA
Figure 2.3: An example of heatmap along with the dendrogram of genes and arrays

Figure 2.4: An illustration of MAplot
samples. In earlier days simple fold change was used to test the differential expression of genes. The procedure was to evaluate the log ratio between two conditions and to consider all genes that differ by more than an arbitrary cut-off value to be differentially expressed. It is known to be unreliable because statistical variability was not taken into account (Chen et al., 1997). Fold change method is subject to bias if the data have not been properly normalized. Tanaka et al., (2000) illustrates the danger of false positive and false negative when looking strictly to fold change.

2.4.2 $t$-test

The $t$-test is a simple, statistical method for detecting differentially expressed genes. When there are replicated samples under each condition the straight forward method is to adopt the traditional two sample $t$-test (Devore and Peck, 1997). Here we summarize the procedure as follows: Let $X_1$ and $X_2$ be two independent gene expression data under two conditions for a particular gene, the two sample $t$-statistic is computed as

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $s^2$, $n_1$ and $n_2$ are the pooled sample variance and the number of observations in each condition respectively and $s^2$ is computed as

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

where $s_1^2$ and $s_2^2$ are the sample variances.

It is known that $t$ follows approximately a Student distribution, with $n_1 + n_2 - 2$ digress of freedom. Because $t$-test utilize the variance of the samples it has potential of addressing some of the shortcomings of fold change approach. However, the small sample sizes in microarray studies can affect the variance estimates.
2.4.3 \( t \)-test (unequal variance)

Student \( t \)-test assumes that features within each group have similar variances. But this is rarely seen in microarray experiments. Hence the Welch (Satterwaite’s) \( t \)-test method developed for independent samples with unequal variances would be much more appropriate. This test is based on the difference between sample means and assumes normal distribution for the samples but allows different variances of samples. Welch method proposes an elaborate correction for degrees of freedom under unequal variances of samples as in Eq. 2.5.

\[
\nu = \left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2 / \left( \frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)} \right)
\]  

(2.5)

where \( s_1^2, s_2^2, n_1 \) and \( n_2 \) are the sample variances and the number of observations in each category (Welch, 1938).

2.4.4 Modifications of \( t \)-test

To address the shortcomings of the \( t \)-test several modifications of the form in Eq. 2.6, have been applied.

\[
t^* = \frac{\bar{M}}{(s + a)/\sqrt{n}}
\]  

(2.6)

With this modification, genes with small fold changes will not be selected as significant. Efron et al., (2001) used a percentile of the distribution of sample standard deviations as \( a \). The SAM \( t \)-test (S-test) estimated \( a \) from all the individual gene variances (Tusher et al., 2001). Broberg (2003) considered the two sample problem and proposed a computationally intensive method of determining the \( a \) by minimizing a combination of estimated false positive and false negative rates over a grid of significance levels and variances.

The regularized \( t \)-test combines information from gene-specific and global average variance estimates by using a weighted average of the two as the denominator.
for a gene-specific $t$-test. The idea of using a $t$-statistic with a Bayesian adjusted denominator was also proposed by Baldi and Long (2001), who developed the useful cyber T program. The statistic is of the form in Eq. 3.3.

$$t^* = \frac{\bar{M}}{s/\sqrt{n}}$$  \hspace{1cm} (2.7)

where

$$s^2 = \frac{\nu_0 \sigma_0^2 + (n-1)s^2}{\nu_0 + n - 2}$$

is the regularized standard deviation. $\nu_0$ is the strength of the prior and $\sigma_0^2$ is the background variance estimated from all genes or a set of subset genes.

All these methods are based on using the two sample $t$ test or its minor variation, differ in how we associate a significance level to the corresponding statistic, leading to possibly large difference in resulting significance levels and number of genes selected. The $p$-value or observed significance level $p$ is the probability of getting a test statistic as or more extreme than the observed one, under the null hypothesis $H_0$ of no differential expression.

2.4.5 Regression modelling approach

Thomas et al., (2001) proposed a regression modelling approach which followed basic idea of Efron et al., (2001). The idea is to model the expression level $y_{gi}$ of gene $g$ ($g = 1, 2...p$) in the array $i$ ($i = 1, 2...n_1, n_1 + 1, ..., n_2$) as follows. Suppose the first $n_1$ and last $n_2$ arrays are obtained under the two conditions respectively, the general statistical model is represented as Eq. 2.8.

$$y_{gi} = \alpha_i + b_gx_i + \epsilon_{gi}$$  \hspace{1cm} (2.8)

where $x_i = 1$, for $i = 1, 2..., n_1$ and $x_i = 0$, for $i = n_1 + 1, ..., n_2$.  

27
The $\epsilon_{gi}$ are the random errors with mean 0. Hence, $b_g$ represents the difference of expression levels of gene $g$ across two conditions. The test for differential expression thus becomes testing for the null hypothesis in $H_0 : b_g = 0$ against $H_1 : b_g \neq 0$.

The $\alpha_i$ and $b_g$ estimated using weighted least square approach and estimated the variance of $\hat{b}_g$ using the robust or sandwich variance estimator. The test statistic in Eq. 2.9 is used with reference to the asymptotically normal distribution.

$$Z_g = \frac{\hat{b}_g}{\sqrt{\text{Var}(\hat{b}_g)}}$$ (2.9)

The final inference to this modelling approach is based on large sample approximations and may be a serious limitation of this approach in microarray studies where the sample sizes are too small.

### 2.4.6 Alternative approaches

The Wilcoxon rank sum test (equivalent to Mann–Whitney test) for two groups or Kruskal-Wallis test for two or more groups has also been used as an alternative method in testing differential expression which can be applied especially when the data do not follow a normal distribution. Critical importance to possibly identify a small subset of genes that best discriminate between tissues under different conditions. The $B$ statistic in Eq.2.10 proposed by Lonnstedt and Speed (2002) is the log posterior odds ratio of differential expression versus non-differential expression.

$$B = \log \frac{\Pr \{ \text{DE} \}}{\Pr \{ \text{not DE} \}}$$ (2.10)

It allows for gene specific variances but it also combines information across many genes using an empirical Bayes approach and thus should be more stable than the $t$-statistic and it is equivalent to moderated $t$-statistics in terms of ranking of genes. The moderated $t$-statistic is shown to follow a $t$-distribution with augmented degrees
of freedom. The moderated \( t \) inferential approach extends to accommodate tests involving two or more contrasts through the use of moderated \( F \)-statistics.

### 2.4.7 Analysis of variance (ANOVA)

The experimental design issues involved in two-color cDNA microarray experiments have recently been receiving more attention. Kerr et al., were first to propose the study of gene expression data using analysis of variance (ANOVA) models (Kerr et al., (2000, 2001a,b)). These models perform both normalization and identification of differentially expressed genes. Kerr et al., (2000) applied ideas from optimal experimental designs to suggest efficient designs for the some of the common microarray experiments. They explored the connection between the microarray designs and classical block design and gave family of ANOVA models as a guide to choose a design. The works of Churchill et al., and Yang et al., have provided comprehensive insights into two-color cDNA microarray experimental design Churchill, (2002); Yang and Speed, (2002, 2003). Yang and Speed, (2002) proposed an average variance approach to choose an experimental design. Churchill, (2002) raised several principles involved in two-color cDNA microarray experimental design.

A typical ANOVA model is in Eq. 2.11:

\[
y_{ijkg} = \mu + A_i + D_j + AD_{ij} + G_g + AG_{ig} + VG_{kg} + DG_{jg} + \epsilon_{ijkg} \tag{2.11}
\]

where \( y_{ijkg} \) is the measured intensity from array \( i \), dye \( j \), variety \( k \) and gene \( g \) on appropriate scale (typically log scale). They called variety to refer the mRNA samples under the study. The varieties may be treatment and control samples, cancer and normal cells or different time points of a biological process. In this model \( \mu \) refers to the overall mean and the terms \( A, D \) and \( AD \) account for all aspects that are not gene specific. The gene effect \( G_g \) capture the average levels of expression of genes and the array-by-gene interaction \( (AG_{ig}) \) accounts for the difference due
to varying sizes of spots on arrays. The dye-by-gene interaction (DG$_{gj}$) represent gene specific dye effects. None of these effects are of biological interest and helps in normalization of the data for ancillary source of variation. The effect of interest is the interaction between the gene and varieties (VG$_{kg}$), differences among these variety by gene interaction accounts for the relative gene expression. For example to estimate the relative gene expression of gene g in varieties 1 and 2, we should estimate (VG$_{1g}$)−(VG$_{2g}$). The error terms $\epsilon_{ijkg}$ are assumed to be independent with mean 0 and variance $\sigma^2$.

The ANOVA has several advantages, first it provides explicit quantitative term for each factor considered in the noise, thus helps in assessing the array as well as dye effects and normalization becomes unnecessary. Wolfinger et al., (2001) considered two interconnected ANOVA model where the residual from the usual ANOVA model which represents the normalized values were fitted separately to a gene model to make inferences of variability. The normalization model is as in Eq. 2.12.

$$y_{gij} = \mu + T_i + A_j + T_A_{ij} + \epsilon_{gij} \quad (2.12)$$

where $\mu$ represents overall mean value, $T$ is the main effect for treatments, $A$ is the main effect for arrays, $T_A$ is the interaction effect of arrays and treatments, and $\epsilon$ is stochastic error. Let $r_{gij}$ denotes the residual from this model, then the gene model under consideration is as in Eq. 2.13.

$$r_{gij} = G_g + GT_{gi} + GA_{gj} + \gamma_{gij} \quad (2.13)$$

where all effects are indexed by $g$ and are assumed to serve the similar role as of normalization model but at the gene level. The effects $A_j$, $T_A_{ij}$, $\epsilon_{gij}$, $GA_{gj}$ and $\gamma_{gij}$ are all assumed to be normally distributed random variables with zero mean and variance components $\sigma^2_A$, $\sigma^2_{TA}$, $\sigma^2_\epsilon$, $\sigma^2_{GA}$ and $\sigma^2_\gamma$ respectively.
Hypothesis testing in ANOVA model involves comparison of null model of no differential expression and an alternative model with differential expression. *F*-statistics are computed for individual genes from the residual sum of squares from fitting of the two models. Thus *F*-statistic is computed as in Eq. 2.14.

\[
F_1 = \frac{(r_{ss0} - r_{ss1})/(df_0 - df_1)}{r_{ss1}/df_1} \tag{2.14}
\]

Other *F*-like statistics (\(F_2\) and \(F_3\)) defined by (Cui and Churchill, 2003), borrow information from other genes for estimating the variance components. \(F_3\) uses the pooled variance estimator \(\sigma^2_{pool}\) for each variance component. For balanced designs, \(\sigma^2_{pool}\) is an average across genes of the individual variance estimates. \(F_2\) uses the average of individual gene variance and pooled variance for each component. Thus \(F_2\) and \(F_3\) are estimated as

\[
F_2 = \frac{(r_{ss0} - r_{ss1})/(df_0 - df_1)}{(r_{ss1}/df_1 + \sigma^2_{pool})/2} \tag{2.15}
\]

\[
F_3 = \frac{(r_{ss0} - r_{ss1})/(df_0 - df_1)}{\sigma^2_{pool}} \tag{2.16}
\]

where \(r_{ss0}, r_{ss1}\), and \(df_0, df_1\) are the residual sum of squares and degrees of freedom for the null and alternative models respectively.

The mixed model ANOVA treats some of the factors in the experimental design as random samples from a population and there are multiple levels of variance (biological, array, spot and residual). Constructing an appropriate *F*-statistics using the mixed model is tricky (Littell et al., 1996)). One can also apply random effects models which use BLUP (best linear unbiased prediction) for the estimation of the gene expression effects (Henderson, 1984).
2.4.8 Multiple Hypothesis Testing

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 \( m \) can be considerably large. 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 (Golub et al. (1999); Tusher et al., (2001); Dudoit et al., (2002)).

Type I and Type II error rates

In any testing situation, two types of errors can be committed: a false positive, or Type I error, is committed by rejecting a true null hypothesis, and a false negative, or Type II error, is committed when the test procedure fails to reject a false null hypothesis. The situation can be summarized as shown in Table 2.1 below, where the number of Type I errors is \( V_n \) and the number of Type II errors is \( U_n \). The numbers \( h_0 \) and \( h_1 \) of true and false null hypotheses respectively and are unknown parameters. The number of rejected hypotheses \( R_n \) is an observable random variable, and the entries in the body of the table, \( U_n, h_1 - U_n, V_n, \) and \( h_0 - V_n \), are unobservable random variables. Ideally, one would like to simultaneously minimize both the chances of committing a Type I error and a Type II error. Unfortunately, this is not feasible and one seeks a trade-off between the two types of errors. A standard approach is to specify an acceptable level \( \alpha \) for the Type I error rate and derive testing
Chapter 2. Review of Literature

Table 2.1: Type I and Type II errors in multiple hypothesis testing.

<table>
<thead>
<tr>
<th>Null hypotheses</th>
<th>Not rejected</th>
<th>Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>True (non-diff.genes)</td>
<td>$h_0 - V_n$</td>
<td>$V_n$ (Type I errors)</td>
</tr>
<tr>
<td>False (diff.genes)</td>
<td>$U_n$</td>
<td>$h_1 - U_n$ (Type II errors)</td>
</tr>
</tbody>
</table>

procedures, i.e., rejection regions, that aim to minimize the Type II error rate, i.e., maximize power, within the class of tests with Type I error rate at most $\alpha$.

The commonly used Type I error rates in multiple hypotheses testing are:

- Per comparison error rate (PCER): the expected value of the number of Type I errors over the number of hypotheses,

$$PCER = \frac{E(V_n)}{m}$$

- Per-family error rate (PFER): the expected number of Type I errors,

$$PFER = E(V_n)$$

- Family-wise error rate: the probability of at least one Type I error

$$FWER = \Pr(V_n = 1)$$

- False discovery rate (FDR): the expected proportion of Type I errors among the rejected hypotheses
\[ \text{FDR} = E \left( \frac{V_n}{R_n} ; R_n > 0 \right) = E \left( \frac{V_n}{R_n} | R_n > 0 \right) \times \text{Pr}(R_n > 0) \]

- Positive false discovery rate (pFDR): the rate that discoveries are false

\[ \text{pFDR} = E \left( \frac{V_n}{R_n} | R_n > 0 \right) \]

**Family-wise error rate**

The family-wise error rate (FWER) is the probability of rejecting at least one true null hypothesis, i.e., the probability of making at least one Type I error. Single-step procedures for strong control of the FWER at level, the Bonferroni procedure rejects any hypothesis \( H_i \) with p-value less than or equal to \( \alpha / m \). The corresponding Bonferroni single-step adjusted p-values (\( \tilde{p} \)) are given by

\[ \tilde{p}_i = \min(m p_i, 1) \]  

(2.17)

Closely related to the Bonferroni procedure is the Sidak procedure, which is exact for protecting the FWER when the raw p-values are independently and uniformly distributed over \([0,1]\) (Sidak, 1967). The Sidak single-step adjusted p-values are given by

\[ \tilde{p}_i = 1 - (1 - p_i)^m \]  

(2.18)

A more general and less conservative definition of adjusted p-values, which takes into account the dependence structure between variables, is proposed by Westfall and Young (1993). The Westfall and Young single-step minP adjusted p-values are defined by

\[ \tilde{p}_i = \text{Pr} \left( \min_{k=1,\ldots,m} p_k \leq p_i | H_0 \right) \]  

(2.19)

where \( H_0 \) denotes the complete null hypothesis.
While single-step procedures are simple to implement, they tend to be conservative for control of the FWER. Improvement in power, while preserving strong control of the FWER, may be achieved by step-down procedures. Let $p_{r1}, p_{r2}, \ldots, p_{rm}$ denote the ordered raw $p$-values. For control of the FWER at level $\alpha$, the Holm’s step down procedure is given by Holm (1979) as

$$\tilde{p}_i = \max_{k=1,\ldots,i} \left\{ \min \left( (m - k + 1)p_{rk}, 1 \right) \right\}$$

Holm’s procedure is less conservative than the standard Bonferroni procedure, which would multiply the $p$-values by $m$ at each step. The Westfall & Young step-down minP adjusted $p$-values are defined by

$$\tilde{p}_i = \max_{k=1,\ldots,i} \left\{ \Pr(\min_{l=k,\ldots,m} p_{rl} \leq p_{rk}|H_0) \right\}$$

Similar method (maxT) under the assumption that the statistics $T_g$ are equally distributed under the null hypothesis and computed as above by replacing $\tilde{p}_i$ by $|T_g|$ and min by max and it is computationally less intensive (Westfall and Young, 1993).

False Discovery Rate

FDR is defined to be the expected value of the ratio of the number of incorrectly rejected hypotheses and the total of number of rejected hypotheses. Controlling FDR proves more powerful than FWER and has become increasingly adopted for genomic studies. Assume a usual $p$-value is available for each hypothesis corresponding to individual genes, based on the $p$-values of the hypotheses, Benjamini and Hochberg (1995) provided a multiple testing procedure that guarantees the FDR to be less than or equal to a prefixed value $q$. The procedure rejects $H_0$ for all genes
for which $p_i \leq p_k$, where

$$k = \max_{i=1,2,...,n} \left\{ i : p_i \leq q \times \frac{i}{n} \right\}$$  \hspace{1cm} (2.22)

where $p_i, i = 1, 2,...,n$, are the $p$ values sorted in ascending order with $p_0 = 0$.

Benjamini and Yekutieli (2001) proposed a simple conservative modification of the original procedure which controls FDR under general dependence. Many related measures such as the positive FDR (pFDR) made popular by Storey (Storey and Tibshirani (2001); Storey (2001, 2002); Storey and Tibshirani (2003)), local FDR (Efron et al., (2001, 2002)) and conditional FDR (Tsai et al., (2003)) have been suggested in the literature. Storey (2002) proposed a mixture model setup for evaluating or estimating pFDR that seems convenient and appropriate in many multiple testing situations. In his framework, the test statistics are supposed to be independent and identically distributed. Each null hypothesis has a fixed probability, $p_0$, of being true. Thus, the number of true null hypotheses, $m_0$, is taken to be a random variable distributed as binomial $(m;p_0)$. Also, marginal $p$-value distribution, $F$, is then a mixture of the uniform distribution (distribution when the null hypothesis is true) and an alternative distribution $F_1$ (distribution when the alternative hypothesis is true). Several papers connect empirical Bayes methods with false discovery rates.

One of the key issues in estimating FDR is the assumption regarding the underlying null distribution. The Significance Analysis of Microarrays (SAM) method uses a full permutation strategy, sampling across all genes and conditions to generate such a null distribution. Also, the popular software SAM computes false discovery rates from a frequentist viewpoint. Furthermore, the use of a mixture of distributions under the null hypothesis of no differential expression and under the alternative hypothesis of differential expression accounts for a certain within-gene dependence. It is also possible to account for multiple testing issues by using such models. Re-
cently methods are developed for improving estimation of false discovery rate by assuming empirical distribution of \( p \)-values (Pounds et al., (2003, 2004)).

### 2.4.9 Bayesian variable selection

Many hypothesis testing problems in microarray rely on the strong distributional assumptions on the data intensities (Newton et al., (2001), Chen et al., (1997), West et al., (2000)). The literature on microarray data is mainly based on two distributions: the log-normal and the gamma distributions, that often appear to be effective when used in a Bayesian hierarchical framework (Newton et al., (2003), Kendziorski et al., (2003)). In the case of empirical Bayes studies, inference is usually made on some quantities related to the posterior distribution of the parameter of interest, or of a certain type of hypothesis. West et al., (2000) considered an approach based on probit model and linear regression for characterizing differential gene expression. Another approach is to model the gene expression data through binary probit model for classification and use Bayesian variable selection methodology to select important genes (Brown et al., (1998)). Lee et al., (2003) proposed a hierarchical Bayesian model and employed latent variables to specialize the model to a regression setting and applied variable selection to select differentially expressed genes. Bae and Mallick, (2004) proposed an alternative method to perform Bayesian gene selection using a two-level hierarchical Bayesian model. Another method which can simultaneously perform gene selection and clustering in high dimensional data has been proposed in microarray data analysis (Tadesse et al., (2005)).

Bayesian models with correlated priors were proposed to model gene expression and to classify between normal and tumor tissues (Ibrahim et al., (2002)). First, empirical Bayes methods have proven to be very efficient, particularly in the situation where the number of observations available for each test is small, which is often the case in cDNA microarray studies. The main advantage of this type of approach over the traditional \( p \)-value based methods, is that it allows a gene-specific
inference, through the use of posterior probabilities, without the need of estimating a set of parameters for each gene. Regardless of the method used to choose the parameters of the model (full Bayes or empirical Bayes), posterior probabilities allow a re-adjustment of the prior information assumed, using the actual observations. As a consequence, posterior inference is in essence less sensitive to extreme values and is less model-dependent, compared to frequentist inference. The parameters of the model are typically estimated using the data for all the genes (in an empirical Bayes manner), allowing for a certain sort of dependence between genes.

2.5 Clustering and Classification

The main types of statistical problems associated with classification of gene expression data are

1. Identification of new classes using gene expression data, cluster analysis / unsupervised learning
2. Classification of malignancies into known classes, discriminant analysis / supervised learning
3. Identification of marker genes that characterize the different classes variable selection

2.5.1 Clustering Algorithms

Beyond identification of differentially expressed genes, clustering of genes from multiple experiments into groups with similar expression patterns is required for further function annotation and diagnostic classification. Clustering can be applied to rows (genes) and/or columns (samples/arrays) of an expression data matrix. Genes clustered in the same group share similar expression profile, which give clues that the unknown genes may have functions or pathways of the respective groups they cluster in. There are many different ways to express and formulate the clustering prob-
lem, and as a consequence, the obtained results and its interpretations depend strongly on the way the clustering problem was originally formulated. For example, the clusters or groups that are identified may be exclusive, so that every instance belongs to only one group. Or they may be overlapping, meaning that one instance may fall into several clusters. Or they may be probabilistic, whereby an instance belongs to each group depending on a certain assigned probability. Sometimes these may be hierarchical, such that there is a crude division of the instances into groups at a high level that is further refined into finer levels. Furthermore, different formulations lead to different algorithms to solve. If we also consider all the variations of each different algorithm proposed to solve each different formulation, we end up with a very large family of clustering algorithms.

A wide range of different methods have been proposed for the analysis of gene expression data including hierarchical clustering, self-organizing maps, and k-means approaches. Many of the proposed algorithms have been reported to be successful but no single algorithm has emerged as a method of choice. Most of the algorithms are based on heuristic methods, and the issues of determining the correct number of clusters and the choice of best algorithm have yet to be solved. Although in the literature there are as many different classifications of clustering algorithms as the number of algorithms itself, there is one simple classification that allows essentially splitting them into the following two main classes:

- Parametric Clustering
- Non-Parametric Clustering

2.5.2 Parametric Clustering

In general, parametric methods attempt to minimize a cost function or an optimality criterion which associates a cost to each instance cluster assignment. The goal of this kind of algorithm is to solve an optimization problem to satisfy the optimality
criterion imposed by the model, which often means minimizing the cost function. This type of method usually includes some assumptions about the underlying data structure.

The following are representative examples.

The Gaussian Mixture model

In this approach, the data are viewed as coming from a mixture of Gaussian distributions, each representing a different cluster. As stated in Fraley and Raftery, (1998) the probability density function (p.d.f) of an observation x in the finite mixture form is:

\[
p(x; \theta) = \sum_{i=1}^{K} \tau_i p_i(x; \theta) = \sum_{i=1}^{K} \tau_i p(x/i; \theta)
\]

\[
= \sum_{i=1}^{K} \tau_i \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu_i)^T (\Sigma)^{-1} (x - \mu_i) \right)
\]

(2.23)

(2.24)

where \(\tau\) is the mixing parameter, \(p(x; \theta)\) is the p.d.f corresponding to the distribution. In case of multivariate Gaussian components, \(\theta\) consists of the elements of the mean vector \(\mu\) and \(\sigma\). The vector of unknowns \((\tau, \theta)\) belongs to some parameter space and is estimated using the Expectation Maximization (EM) algorithm. The negative log-likelihood for the data set is given by

\[
E = -\ln L = - \sum_{j=1}^{m} \ln p(x_j) = - \sum_{j=1}^{m} \left( \sum_{i=1}^{K} \tau_i p(x_j/i) \right)
\]

(2.25)

which can be an error function that will be minimized. Here the algorithm using EM is summarized as follows.

1. Start with \((\mu_i^0, \sigma_i^0, \tau_i^0)\)
2. Having \((\mu_i^t, \sigma_i^t, \tau_i^t)\), the next iteration is obtained as

\[
\mu_i^{t+1} = \frac{\sum_{j=1}^{m} p^t(i/x_j)x_j}{\sum_{j=1}^{m} p^t(i/x_j)}
\]
$$\mu^t_i + 1 = \frac{\sum_{j=1}^m p^t(i/x_j) \| x_j - \mu^t_i \|^2}{\sum_{j=1}^m p^t(i/x_j)}$$

$$\tau^t_i + 1 = \frac{1}{m} \sum_{j=1}^m p^t(i/x_j)$$

where

$$p^t(i/x_j) = \frac{p^t(x_j/i) \tau^t_i(i)}{p^t(x_j)}$$

3. Calculate the change in the error function:

$$\Delta^t + 1 = E^{t+1} - E^t = -\sum_{j=1}^m \ln \frac{p^{t+1}(x_j)}{p^t(x_j)}$$

4. if $\Delta < tol$ stop, otherwise go to step 2.

Fuzzy c means clustering

Partitioning methods aim to find the best partition of the genes / samples into $C$ groups in such a way that one criterion is optimized. Here we used FCM algorithm Bezdek, (1981) for partitioning the samples into two groups based on the selected genes. The fuzzy clustering assigns different degrees of membership to each sample to a particular group and the membership values lies between 0 (not in the group) and 1 (completely in the group). The criterion which is minimized iteratively is the weighted within group sum of squared distance $d(x_j, c_i)$ with weights as the membership values $u_{ij}$, and is given by

$$J_m(C, M) = \sum_{i=1}^C \sum_{j=1}^M (u_{ij})^m d^2(x_j, c_i), \tag{2.26}$$

where $C$ is the total number of clusters and $M$ is the total number of samples to be clustered, $m$ is the fuzziness parameter and $d(x_j, c_i)$ is the distance of the sample $x_j$ to the centroid $c_i$ and $u_{ij}$ is the membership value. Our aim is to obtain the centroid vectors and corresponding membership values. Here the algorithm is summarized.
as follows.

1. Choose primary centroids \(c_i\) arbitrarily.

2. Compute the degrees of membership values

\[
u_{ij} = \frac{\left( \frac{1}{d^2(x_j, c_i)} \right)^{\frac{1}{m-1}}}{\sum_{k=1}^{C} \left( \frac{1}{d^2(x_j, c_k)} \right)^{\frac{1}{m-1}}} ;
\]

where, \(m\) is the fuzziness parameter and here we set it to be 1.75.

3. Compute the new centroids \(\hat{c}_i\) where

\[
\hat{c}_i = \frac{\sum_{j=1}^{M} (u_{ij})^m x_j}{\sum_{j=1}^{M} (u_{ij})^m}
\]

and update the memberships, \(u_{ij}\) to \(\hat{u}_{ij}\) according to step 2.

4. If \(\max \| u_{ij} - \hat{u}_{ij} \| < \epsilon\) (tolerance limit) then we stop, otherwise go to step 3.

**K-means and K-median**

The K-means algorithm, probably the first one of the clustering algorithms proposed, is based on a very simple idea: Given a set of initial clusters, assign each point to one of them, then each cluster center is replaced by the mean point on the respective cluster. These two simple steps are repeated until convergence. A point is assigned to the cluster which is close in Euclidean distance to the point. Although K-means has the great advantage of being easy to implement, it has two big drawbacks. First, it can be really slow since in each step the distance between each point to each cluster has to be calculated, which can be really expensive in the presence of a large dataset. Second, this method is really sensitive to the initial clusters, however, in recent years, this problem has been addressed with some degree of success.
2.5.3 Non-Parametric Clustering

Two good representative examples of the non-parametric approach to clustering are the agglomerative and divisive algorithms, also called hierarchical algorithms, which produce dendrograms. Both methods are based on the measures of the dissimilarities among the current cluster set in each iteration. Agglomerative algorithms merge some of the clusters, depending on how similar they are, and divisive algorithm splits them. A dendrogram is simply a tree that shows which cluster were agglomerated in each step. It can be easily broken at selected links to obtain cluster or groups of desired cardinality or radius. This number of cluster or groups can also be determined as a function of some merging threshold. The idea is that with a threshold of zero, the number of clusters is equal of the number of data points, and with a high threshold the data is partitioned in just one single cluster. Since the dendrogram is just a tree, its structure representation is also easy to generate and to store. An important advantage of non-parametric clustering methods is that they do not make any assumptions about the underlying data distribution. Furthermore, they do not require a explicit representation of the data in a Euclidean form. They only require a matrix with the pair wise similarities based on a predefined distance. The Euclidean distance is usually used for individual points. There are no known criteria of which clustering distance should be used, and it seems to depend strongly on the dataset. Among the most used variations of the hierarchical clustering based on different distance measures are

1. Average linkage clustering: The dissimilarity between clusters is calculated using average values. The average distance is calculated from the distance between each point in a cluster and all other points in another cluster. The two clusters with the lowest average distance are joined together to form the new cluster.
2. **Centroid linkage clustering**: This variation uses the group centroid as the average. The centroid is defined as the center of a cloud of points.

3. **Complete linkage clustering (Maximum or Furthest-Neighbor Method)**: The dissimilarity between 2 groups is equal to the greatest dissimilarity between a member of cluster \( i \) and a member of cluster \( j \). This method tends to produce very tight clusters of similar cases.

4. **Single linkage clustering (Minimum or Nearest-Neighbor Method)**: The dissimilarity between 2 clusters is the minimum dissimilarity between members of the two clusters. This method produces long chains which form loose, straggly clusters.

5. **Ward's Method**: Cluster membership is assigned by calculating the total sum of squared deviations from the mean of a cluster. The criterion for fusion is that it should produce the smallest possible increase in the error sum of squares.

**References**


References


gene expression with a semiparametric hierarchical mixture method. Technical Report,

atives in microarray studies by approximating and partitioning the empirical distribution of

20: 1737-1745.

418-427.

R Development Core Team. R (2005). A Language and Environment for Statistical Comput-


York.


Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal dis-

expression in microarray experiments. Statistical Applications in Genetics and Molecular
Biology. 3(1): 3.

with applications to DNA microarrays. Technical Report 28, Department of Statistics,
Stanford University.

Storey, J. D. (2001). The positive false discovery rate: A Bayesian interpretation and the
q-value. Technical Report 12, Department of Statistics, Stanford University.

Storey, J. D. (2002). A direct approach to false discovery rates. Journal of the Royal Statis-
tical Society, Series B 64: 479-498.

Storey, J. D. and Tibshirani, R. (2003). Statistical methods for identifying differentially ex-


Welch BL. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 1938; 29:350-362.


